UNDERSTANDING HUMANS AND OBJECTS BASED ON CONTEXT

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Henry Shu

August 2013

ABSTRACT

Contextual information, such as positions, of each entity of interest, such as human, in a photo or video can provide much information and knowledge about the content of the photo or video in question. In this thesis work, we investigate the benefit of analyzing positions and context under various novel settings. Furthermore, we explore ways of extracting the 3D position information about humans from a single photo. Our work indicates that incorporating the spatio-temporal constraints for moving cars with their natural speed upper bound can greatly improve the driving history reconstruction of the car in question. In addition, knowing the positions of each human relative to one another in a photo is shown to improve the prediction of the age and gender of each. Furthermore, the relative positions of the faces in a group shot can reveal the type of social event under which this shot was taken. We also show how to extract approximate 3D position information about humans based only on a single 2D photo. In conclusion, this thesis affirms the benefits of analyzing positions of objects of interests in various novel settings and opens a vista for further research in the area.

This document is dedicated to all Cornell graduate students.

ACKNOWLEDGEMENTS

Special thanks to Professor Tsuhan Chen and Andrew Gallagher, for their continuous support for this research.

TABLE OF CONTENTS

Dedication 4
Acknowledgements 5
Table of Contents 6
List of Tables 8
List of Figures 9

1 Wide Area Video Surveillance with Spatial-Temporal Constraints 1
1.1 Introduction 1
1.2 Proposed Method 4
1.2.1 Formulation 4
1.2.2 Algorithm 6
1.3 Interactive User Feedback 9
1.4 Experiments 10
1.5 Conclusion 13

2 Modeling Proximal Dependency in Consumer Photos 14
2.1 Introduction 14
2.2 Approach 17
2.3 Experiments 21
2.3.1 Gender Classification 21
2.3.2 Age Classification 24
2.3.3 User Feedback 25
2.4 Human Studies 26
2.5 Conclusion 27

3 Relative Depth Estimation in A Group Photo 34
3.1 Introduction 34
3.2 Related Work 36
3.3 Data Collection 38
3.4 Body Contact and z-Coordinate 39
3.5 Algorithms 40
3.5.1 Feature Extraction 42
3.5.2 Classification 43
3.5.3 z-Coordinate Assignment 43
3.6 Experiments 44
3.7 Conclusion 46

4 Face-Graph Matching for Classifying Groups of People 47
4.1 Introduction 47
4.2 Ground Truth and Data Collection 50
4.3 Method 50
4.3.1 Bipartite Matching 51

4.3.2 Face Number Discrepancy 52
4.4 Experiments 53
4.4.1 Main 53
4.4.2 Horizontal Symmetry 55
4.4.3 Effect of Training Size 55
4.5 Conclusion 56

Bibliography **58**

LIST OF TABLES

1.1 Performance improvement of our methods. 12
1.2 Path reconstruction. The top, middle, and bottom rows correspond to 9L-0767, 5C-2717, and DI-8676, respectively. 13

2.1 Performance improvement of our methods in gender classification. 21
2.2 Performance improvement of our methods in age classification. . 25

3.1 Pairwise performance results for depth order. 44
3.2 Pairwise performance results for body contact. 45
3.3 Pairwise performance results for relative distance. (1) is less than 8cm, (2) is between 8cm and 100cm, and (3) is greater than 100cm. 45

LIST OF FIGURES

1.1 (a) License plate matching is a challenging task for vision-based algorithms (Baseline), whose performance can be improved by our method (Proposed), which optimally solves a constrained retrieval problem. (b) Our method can reconstruct the driving path very accurately without using or learning any location transition probabilities. 2

1.2 Shown are 13 video frames taken from 3 cameras (red, blue, and green). A purple link connects two video frames that do not obey the speed limit constraint. For simplicity, only speed limit violations relative to f_{12} are drawn. In reality, N is usually > 1000. . . . 6

1.3 (a)(b) Treating the selected video frames as ranked (by similarity scores) retrievals, the precision-recall curves show the quality of our algorithms. (c) 16 of the unselected frames returned for human feedback for query 5C-2717. They are automatically suggested by our algorithm. The top (bottom) number is camera (timestamp). 11

2.1 Without appearances, position cues for estimating gender and age are weak (a) - (e), difficult even for humans. With the extra information that (b) and (d) belong to a two-person photo (f), and that (a), (c), and (e) belong to a three-person photo (g), it becomes easier to estimate that (b) is male, (d) is female, (e) is an adult male, (a) is an adult female, and (c) is a teenager. In this paper, we harness this extra information (h). Incorporating the appearance-based cues (i) in a principled way, we demonstrate that our method can (j) further enhance current appearance-based state-of-the-art. 28

2.2 (a) and (c) use marginal fit potentials, while (b) and (d) use learned potentials. (a) and (b) use no MRF edges, while (c) and (d) use MRF edges. (e) illustrates the idea of our algorithm. . . . 29

2.3 Gender classification result averaged over 20 independent runs. 29

2.4 Except RAND, each data point in (a) - (d) is averaged over 20 independent runs. 30

2.5 Age classification result averaged over 20 independent runs. . . 30

2.6 User feedback experiment. Each bar averaged over 50 independent runs. 31

2.7 (a) Distribution of the number of faces. (b) Gender accuracy with different graphs, averaged over 50 independent runs. 31

2.8 (a) Age cls. performance over training sizes, averaged over 20 independent runs. (b) The comparison between human performance and the proposed algorithm. Results averaged over 10 human respondents for age, and 13 for gender. 32

2.9 Fig. (a) shows the original photo. Fig. (b) shows the picture we provide to the respondent for interpreting the gender. 32
2.10 Graphical model structures including the (a) complete, (b) star, (c) shortest Hamiltonian path, (d) minimum spanning tree, and (e) Delaunay triangulation graphs. (Better viewed in color) . . . 33
2.11 Results of gender (top) and age (bottom) classification using positions only (without any appearance features). The MRF structures from PP are shown. Circles are correct predictions and crosses are wrong predictions. For age, the top number is the ground truth, and the bottom one is the prediction (shown only if different from the ground truth), with the finer 7 age groups. Best viewed in electronic version. 33

3.1 Given an input image (a) our algorithm estimates the z-coordinate of each person, rendered in (b). We model this problem as a joint classification of all the 3 modules. (c) The result is the prediction of the relative distance of each person to the camera (encoded in grayscale) and whether pairs of people are in physical contact (blue edges). The numbers indicate the sorting of the people in their z-coordinates. 35
3.2 (a) Sample Mechanical Turk HIT interface. (b) These distances are inconsistent with $A <_z B <_z C$. Our model addresses this problem by using a z-coordinate assignment algorithm that guarantees global consistency. 38
3.3 41
3.4 41
3.5 (a) The original cropped image of a pair of people. (b) The HOG rendering. (c) The sparsity of SVM makes many of the feature points vanish. (d) 5 clusters of the normalized xy positions of pairs of people in our data set. 45

4.1 The label space is (1) Family, (2) Group Field Trip, (3) Sports Team, and (4) Friends Hanging Out 48
4.2 Answers to Fig. 1. The label space is (1) Family, (2) Group Field Trip, (3) Sports Team, and (4) Friends Hanging Out 49
4.3 (a) and (b) are two sample photos showing the face bipartite graph. (c) is the core experiment result. 51
4.4 54

4.5 Shown in each of (a) - (h) are two image pairs. In each pair, the left image is the test query and the right is its most similar image from the training set. The left pair of images is based on POS, and the right pair of images is based on POS+AG, in which the gender and age predictions are shown as well. The ground truth photo types are provided at the bottom of each image. Best viewed in magnification in color. 57

CHAPTER 1

WIDE AREA VIDEO SURVEILLANCE WITH SPATIAL-TEMPORAL CONSTRAINTS

1.1 Introduction

With the large-scale deployment of surveillance cameras along city streets, the streaming of surveillance videos have become useful resources for tracking suspicious vehicles. Unlike video surveillance for small areas where objects can be tracked across cameras based on overlapping regions, cameras deployed in wide areas are not dense enough to enable seamless tracking. Therefore, object tracking algorithms via overlapping fields of view widely adopted for small-area surveillance cannot be utilized. In addition, the cameras deployed in city streets are likely from a variety of manufacturers, models, and imaging characteristics including resolution, color temperature, etc. In contrast to small-area surveillance, these cameras have a much higher degree of heterogeneity that is very expensive to overcome. With other uncontrollable factors from the environment, such as inconsistencies in weather, air quality, visibility, and traffic, it is not surprising that vision-based matching or recognition algorithms alone, such as query matching or OCR, can perform quite poorly in wide-area surveillance.

In the literature, some works have been published to address these challenges for wide-area surveillance. Kettnaker and Zabih [36] were among the first to utilize the constraints on the motion of the objects across cameras, where the path topology and transition probabilities across non-overlapping cameras are assumed to be known. In their work, although the maximum a posteriori solu-

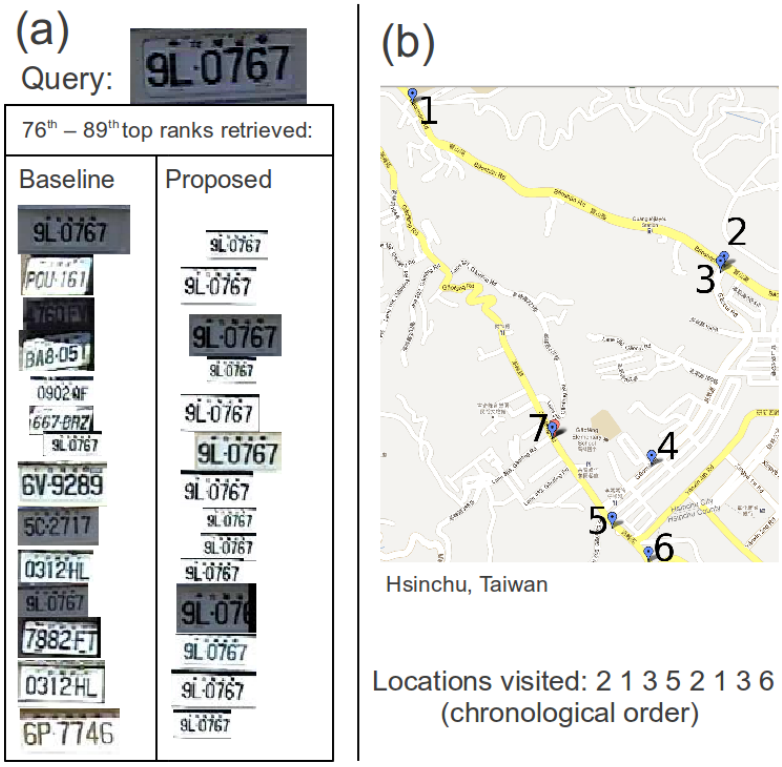


Figure 1.1: (a) License plate matching is a challenging task for vision-based algorithms (Baseline), whose performance can be improved by our method (Proposed), which optimally solves a constrained retrieval problem. (b) Our method can reconstruct the driving path very accurately without using or learning any location transition probabilities.

tion can be efficiently approximated by linear programming, the requirement of knowing the transition probabilities hinders its practical use. Collins et al [12] proposed to use calibrated cameras with a known 3D environment model to track objects across multiple cameras. However, this is only practical for small areas where camera calibration and 3D site model construction can be done without too much effort. In [32], Javed et al alleviated the requirements of camera calibration and known path topology by formulating tracking across multiple non-overlapping cameras as the path cover problem in a directed graph. The goal is to find the hypothesis that maximizes a posteriori probability of the as-

sociation given the observations of individual tracks in each camera. The probability of track association across two cameras is established based on the object appearance modeled by color histogram and the space-time features modeled by location, velocity, and time. To further handle the appearance differences of the same object across different cameras, Javed et al [33] proposed to learn the brightness transfer function (BTF) by assuming that this inter-camera transformation lies in a low-dimensional subspace. While the works [32, 33] do offer improvement over previous methods, their practical applicability can be limited. Firstly, their algorithm runs in time $O(N^{2.5})$, where N, the number of observations, is typically thousands or more. This makes it too slow to run in practice without some approximation schemes [33]. Secondly, they require the learning of camera transition probabilities from data. Oftentimes, there are not sufficient data available for learning. Also, people have different driving behaviors. Indeed, for a suspect vehicle escaping a crime scene, the route taken may well be one that is least likely to happen. What is more, when the cameras are sufficiently far apart from one another, the transition probabilities tend to be uniform, which makes them less useful. In this paper, we seek a method that does away with transition probabilities altogether.

Alternatively, some recent attempts try to cast the problem as a content-based image retrieval (CBIR) problem ([4, 42]). In CBIR, an image is given as the query, and the task is to rank images in the archive according to their similarities to the query. This methodology is readily extensible to relevance feedback, in which a human user helps winnowing the top ranked images as true hits and iteratively guides the computer in ranking the images. It has been shown that such active learning is an efficient way to annotate a large-scale image dataset [15, 14]. In [4], Ali et al proposed a metric learning algorithm based on relevance feedback.

The idea is to find a linear transformation that maximizes (minimizes) the distances between the query and irrelevant (relevant) images. Our work here is also a variant of CBIR. Different from [4], we do not focus on distance metrics or feature spaces. Instead, we seek to optimally select a set of images subject to spatial-temporal constrains.

The crux of our algorithm lies in the observation that the original problem can be decoupled into independent subproblems. This is similar in spirit to Markov chain and other graphical model-based methods, which utilize independence to make inference tractable. In fact, our speed limit constraint (see next section) can be equivalently formulated as a graphical model problem, where an edge represents a speed limit violation between two video frames. However, a typical problem instance has hundreds of thousands of such edges, and our first experimentation using graphical models indicates a typical running time of > 20 minutes. Furthermore, the temporal gap constraint, which enforces a global upper bound on the number of selected video frames that are temporally far apart from each other (see next section), cannot be readily encoded into a graphical model.

1.2 Proposed Method

1.2.1 Formulation

Let f_i, $1 \leq i \leq N$, be a video frame taken from camera c_i at time t_i. Let us order the video frames temporally, so that $t_1 \leq t_2 \leq \ldots \leq t_N$, breaking ties arbitrarily. For each pair of cameras c^1 and c^2, let $d(c^1, c^2)$ be the shortest driving distance

from the surveillance range of camera c^1 to that of c^2. For each i, let s_i be the inclusion cost of selecting f_i, and let z_i be the exclusion cost of not selecting f_i. (See our experiments for how these costs are obtained.) We formulate the wide area surveillance problem as one in which we seek some subset (the selected frames) $S \subseteq \{1, 2, \ldots, N\}$ so that

$$M \equiv \sum_{i \in S} s_i + \sum_{i \notin S} z_i \tag{1.1}$$

is minimized, subject to two constraints as described below.

Speed limit. If $i, j \in S$, with $i < j$, then $d(c_i, c_j)/(t_j - t_i) \leq \gamma$. Here, γ is a parameter that dictates an upper bound for the speed of the vehicle under surveillance. Intuitively, this constraint ensures that the time it takes to travel from c_i to c_j is physically possible. Notationally, write $(i, j) \in E$ whenever f_i and f_j satisfies the *speed limit* constraint.

Temporal gap. If $i, j \in S$ and no k between i and j ($i < k < j$) is in S, then f_i and f_j are said to induce a *temporal gap* if $t_j - t_i \geq \tau$. Here, τ is a parameter that dictates when two selected frames are temporally far apart. The temporal gap constraint requires that the selected frames in S altogether induce no more than K temporal gaps. Note that a temporal-gap pair (i, j) may either satisfy (A) $c_i = c_j$ or (B) $c_i \neq c_j$. If we further restrict the temporal-gap pairs to only come from type (A), then this reduces to a smoothness constraint. It would ensure that the selected video frames come as temporal clusters, which is intuitively satisfying. However, empirically we find that only using type (A) result in too many false positives. The inclusion of type (B) alleviates this problem by indirectly limiting the number of temporal clusters, since temporally consecutive clusters often come from different cameras. Alternatively, it is possible to only use type (A) while directly upper bounding the number of temporal clusters.

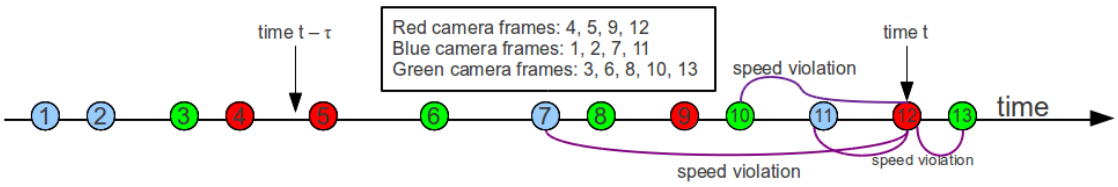


Figure 1.2: Shown are 13 video frames taken from 3 cameras (red, blue, and green). A purple link connects two video frames that do not obey the speed limit constraint. For simplicity, only speed limit violations relative to f_{12} are drawn. In reality, N is usually > 1000.

1.2.2 Algorithm

We now give an efficient algorithm that finds a global optimum to the optimization problem above. First, let us denote by $subp(u, c, k)$ the subproblem in which (C) we assume that there were only u frames $f_1, f_2, \ldots, f_u$ in total, (D) the latest selected frame (not necessarily f_u) is from camera c, and (E) the selected frames S must induce exactly k temporal gaps. Let the optimal cost of $subp(u, c, k)$ be $M(u, c, k)$. Clearly, the original problem is exactly $subp(N, c^*, k^*)$, where $(c^*, k^*) = \arg\min_{(c,k),k \leq K} M(N, c, k)$. Observe that the original problem can be decoupled into subproblems. For example, in Figure 1.2 where $N = 13$, if the latest selected camera is blue, then f_{12} and f_{13} are necessarily unselected, and so we have $M(13, blue, k) = M(11, blue, k) + z_{12} + z_{13}$. In this case, we reuse $subp(11, blue, k)$ (whose optimal solution is not necessarily one that selects f_{11}) to save computation. Our algorithm recursively computes the optimal solution for each $subp(u, \cdot, k)$ from $k = 0$ to K and $u = 1$ to N. In each iteration of u, our algorithm constructs and makes use of several auxiliary data structures defined as follows.

- $T(u, k)$: This is the latest selected frame index of an optimal solution to

$subp(u, c_u, k)$. $T(u, k)$ may possibly $\neq u$.

- $L(u, k)$: Consider an optimal solution to $subp(u, c_u, k)$ requiring that f_u is selected (even if selecting f_u does not actually give a global optimum to $subp(u, c_u, k)$). $L(u, k)$ is the latest selected frame index prior to f_u in this solution.
- $Q(u, k)$: This is the cost of an optimal solution to $subp(u, c_u, k)$ requiring that f_u is selected.

We now give an illustration of the computation of $M(12, \cdot, k)$ from Figure 1.2 by considering the following cases.

Not selecting. In $subp(12, \cdot, k)$, if the latest selected frame is not $c_{12} = red$, then f_{12} is necessarily unselected, and so

$$M(12, blue, k) = M(11, blue, k) + z_{12} \tag{1.2}$$

$$M(12, green, k) = M(10, green, k) + z_{11} + z_{12} \tag{1.3}$$

It is possible that the latest selected frame is $c_{12} = red$, but f_{12} is not selected. In this case, the corresponding cost is $m_0 = M(9, red, k) + z_{10} + z_{11} + z_{12}$.

Selecting without a temporal gap. In $subp(12, \cdot, k)$, if f_{12} is selected without introducing a temporal gap, then the previous selected frame must be one of f_5, f_6, f_8, and f_9. (f_7, f_{10}, and f_{11} are not possible because they have speed limit violation with f_{12}). Therefore, the corresponding cost is

$$\begin{aligned} m_s = \min(&Q(5, k) + z_6 + z_7 + \ldots + z_{11} + s_{12}, \\ &Q(6, k) + z_7 + z_8 + \ldots + z_{11} + s_{12}, \\ &Q(8, k) + z_9 + z_{10} + z_{11} + s_{12}, \\ &Q(9, k) + z_{10} + z_{11} + s_{12}). \end{aligned} \tag{1.4}$$

Let v_s be one of 5, 6, 8, and 9 that achieves the minimum cost m_s above.

Selecting with a temporal gap. In $subp(12, \cdot, k)$, if f_{12} is selected while introducing a temporal gap, then the previous selected frame can be no later than f_4. By subproblem decoupling, the corresponding cost is

$$\begin{aligned} m_g = \min(&M(4, red, k-1) + z_5 + z_6 + \ldots + z_{11} + s_{12}, \\ &M(3, green, k-1) + z_4 + z_5 + \ldots + z_{11} + s_{12}, \\ &M(2, blue, k-1) + z_3 + z_4 + \ldots + z_{11} + s_{12}) \end{aligned} \quad (1.5)$$

Note that, once again, $M(2, blue, k)$ does not necessarily correspond to f_2 being selected. Let v_g be one of 2, 3, and 4 that achieves the minimum cost m_g above.

Auxiliary data structures. Finally, we can set $M(12, red, k) = \min(m_0, m_s, m_g)$. The auxiliary data structures Q, T, and L also need to be computed, since they will be used recursively. By definition of Q, we set $Q(12, k) = \min(m_g, m_s)$. Also, the definition of T implies that $T(12, k) = 12$ only if f_{12} is selected in $subp(12, c_{12}, k)$. Therefore, T is updated as

$$T(12, k) = \begin{cases} T(9, k) & \text{if } m_0 < \min(m_s, m_g), \\ 12 & \text{otherwise.} \end{cases}$$

Here, 9 is the latest frame earlier than f_{12} that is also from camera $c_{12} = red$. Also, in an optimal solution to $subp(12, red, k)$ requiring that f_{12} is selected, the selected frame prior to f_{12} depends on whether f_{12} introduces a temporal gap with it, as follows.

$$L(12, k) = \begin{cases} v_s & \text{if } m_s < m_g, \\ T(v_g, k-1) & \text{otherwise.} \end{cases}$$

Finally, our algorithm iterates in this fashion from $k = 0$ to K and $u = 1$ to N, setting the base cases appropriately. Specifically, $k = 0$ is the base case in which

no video frames are selected. Note that our algorithm does not commit prematurely. Even if $\min(m_s, m_g) < m_0$ during an iteration for u, f_u is not necessarily selected in the final solution. Instead, once M and L have been computed, we reconstruct the final minimum-cost video frame selection S as follows. First, let $(c^*, k^*) = \arg\min_{(c,k),k\leq K} M(N, c, k)$. Then, starting with $u^* = T(u', k^*)$ as the latest selected frame index, where u' is the latest video frame for which $c_{u'} = c^*$, we iteratively trace back all the selected frames using L, decrementing k by 1 whenever we encounter a temporal gap between two consecutively selected frames. Note that this algorithm, which we abbreviate as PP, is completely automatic and requires no human interaction.

1.3 Interactive User Feedback

The algorithm above works without any human intervention. With the availablility of a human user in the loop, we now seek to devise a method in which a small set of video frames are returned to the user for manual feedback (including or excluding). To minimize human effort, the desired property of the feedback mechanism is that only a small set of video frames for feedback are required to further improve the performance of WAVS. Let Y be the (small) set of indexes of video frames for human feedback. Once we obtain the user-provided labels for each $i \in Y$, we can force s_i or z_i to ∞ for these frames and rerun our algorithm to get an updated frame selection set S. The user feedback mechanism then repeats in this manner.

To improve recall, our feedback algorithm computes Y as follows. First, the set S' of unselected frames is computed, where each member $j \in S'$ has the property that the single addition of j to S (which is the optimally selected frames

returned by PP) violates neither of our spatial-temporal constraints. Then, the frame index $j^* \in S'$ with the highest exclusion cost z_{j^*} is selected as a candidate for the human user to provide feedback. Depending on the human feedback, we reset one of s_{j^*} or z_{j^*} to ∞ and rerun PP. The human feedback phase can then repeat in this manner. In Figure 1.3(c), we show 16 feedback candidates suggested by our algorithm during the course of human interaction. Empirically, a good number of frames in Y are true hits. The true positives in Y tend to be visually difficult cases, with low-resolution or occluded license plates.

1.4 Experiments

Here, we describe the experiments we conduct to evaluate our method. We collect our own datasets by having our colleagues driving around the vicinity of Fig. 1.1(b) during peak traffic hours to ensure that many other vehicles are also present. Each driving trip takes a minimum of 20 minutes. The videos from these cameras are later obtained by the local police station. Then, we use the techniques from [34] for license plate detection to pick only those video frames with the presence of any vehicle. We manually label each of these frames as a match or a mismatch with respect to our colleague's vehicles. These constitute the N frames in our algorithm above. Shown in Table 1.2 are three of our colleagues' driving trips.

Given a query input image I, say 9L-0767 as in Figure 1.1(a), we compute for each of the N frames a similarity score h_i that measures how similar frame f_i is to I ($0 \leq h_i \leq 1$, with 0 being the most similar). We use the work of [56] for computing h_i, as dynamic time warping is robust even for a partially occluded license plate. Following [4], our baseline (abbreviated as BL) sorts the N frames

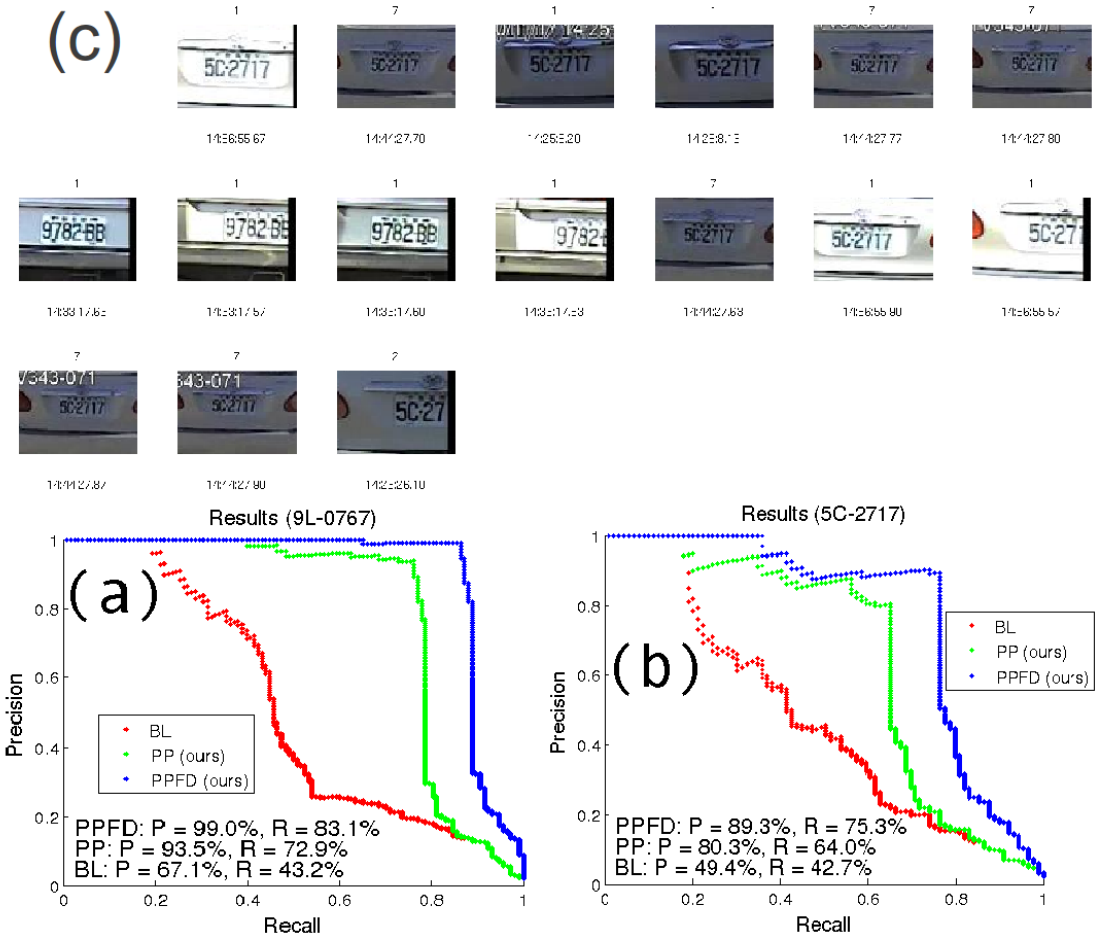


Figure 1.3: (a)(b) Treating the selected video frames as ranked (by similarity scores) retrievals, the precision-recall curves show the quality of our algorithms. (c) 16 of the unselected frames returned for human feedback for query 5C-2717. They are automatically suggested by our algorithm. The top (bottom) number is camera (timestamp).

by the h_i's and returns the top L frames as the selected frames. To make the case favorable for BL, for each input vehicle we pick the particular L so as to maximize the F1 score of BL (see Table 1.1). Of course, no such leniency is applied to evaluate our proposed method. To run PP, we need the inclusion and exclusion costs s_i and z_i, which are computed as $z_i = -\log(1 - \exp(-\lambda h_i))$ and $s_i = \lambda h_i$, where $\lambda > 0$ is a paramter. There is a probabilistic motivation for computing

Plate	Metric	BL	PP	PPFD	Info
9L-0767	Precision	67.1%	**93.5%**	**99.0%**	$N = 6062$
	Recall	43.2%	**72.9%**	**83.1%**	
	F1	52.6%	**81.9%**	**90.4%**	
5C-2717	Precision	49.4%	**80.3%**	**89.3%**	$N = 4596$
	Recall	42.7%	**64.0%**	**75.3%**	
	F1	45.8%	**71.2%**	**81.7%**	
DI-8676	Precision	26.4%	**91.7%**	**97.6%**	$N = 4249$
	Recall	24.6%	**38.6%**	**71.9%**	
	F1	25.5%	**54.3%**	**82.8%**	

Table 1.1: Performance improvement of our methods.

the costs this way. Indeed, if we let p_i be the probability of including f_i and set $p_i = \exp(-\lambda h_i)$, then these costs are precisely the negative log-likelihood. For all of our experiments, we fix $\lambda = 3.8$ and $K = 10$. We abbreviate by PPFD our proposed method with interactive human feedback. For PPFD, we return $|Y| < 20$ frames for the human user to provide feedback, and report the performance after $|Y|$ feedback inputs. Table 1.1 summarizes the performance improvement of our methods over BL. The performance metrics are the video frame-level precision and recall. Both BL and PP use the same similarity scores [56]. However, by globally optimizing the frame selection costs with respect to the spatio-temporal constraints, our methods greatly outperform BL.

We can sort the selected video frames temporally and reconstruct the path the input vehicle has traveled by listing the camera locations visited. This could be useful when, for example, analyzing the escape pattern of a suspect leaving a crime scene. Table 1.2 lists the paths reconstructed using BL and PP. Even without any human feedback, our method PP can reconstruct the paths very accurately.

BL Path	PP Path	True Path
2 1 2 1 2 1 3 5 2 1 3 1 3 1 3 6	2 1 3 5 2 1 3 6	2 1 3 5 2 1 3 6
2 1 3 1 7 5 3 1 3 1 3 6	2 1 3 7 1 3 6	2 1 3 7 5 1 3 6
1 2 1 2 1 2 1 2 1 3 1 2 1 3 1 3 6	2 1 3 2 1 3 6	2 1 3 2 1 3 6

Table 1.2: Path reconstruction. The top, middle, and bottom rows correspond to 9L-0767, 5C-2717, and DI-8676, respectively.

1.5 Conclusion

We have developed a novel method for the problem of target retrieval for wide area video surveillance. Different from existing methods, our algorithm does not require the learning or usage of any transitional probabilities nor the construction of the site model. We formulate the task as a minimum-cost frame selection problem with two spatial-temporal constraints that respect physical reality. Our algorithm can find a global optimum to the optimization problem and greatly outperforms the baseline. Furthermore, it reconstructs the path traveled almost perfectly. Empirically, our algorithm takes under 10 seconds to run in a 1.67GHz laptop, and is thus feasibly deployable on a handheld device.

CHAPTER 2

MODELING PROXIMAL DEPENDENCY IN CONSUMER PHOTOS

2.1 Introduction

When taking photos, people position themselves in a manner that is far from random. The factors that affect the way people position themselves can be physical, cultural, or psychological. Physical factors include such patterns as tall people standing in the back row. Cultural influences include the persons of importance (bride and groom in a wedding photo, or the person of honor in an accolade ceremony) being centered in the middle, or the adults holding the children.

In many of these scenarios, the position of a face relative to a group of neighbors matters. In fact, our studies show that, in the absence of any appearance-based cues, humans rely heavily on the genders and ages of *all* the other faces in a photo to determine those of an unknown one. In this work, we seek to model such dependence in the application of gender and age classification for each face in a photo. That is, in our model, the gender and age likelihood estimate of any face depends on that of any other face in the same photo, and vice versa. We believe that such joint modeling, to a first approximation, can mimic the way humans behave in a photo-shooting setting.

The baseline we use is the pioneering work [20], which used facial position cues to yield gender and age classification results that were better than random. The position cues were also shown to boost the performance of existing appearance-based gender/age classifiers [25, 26]. In [20], the task was formulated as an instance-independent classification problem. That is, for example,

the gender/age classification of each face in photos like Figure 2.1(f)(g) would be treated as five independent instances as Figure 2.1(a)-(e) (after positional mean removal). This formulation works great to pick up and take advantage of such patterns as men tending to stand toward the two sides in a photo, babies tending to be at the bottom of a photo, etc. In this work, we attempt to explicitly model the dependency of the faces in an MRF framework. Our goal is not to compete with existing appearance-based classifiers. Rather, we aim to demonstrate that modeling dependency can further boost the benefits of positional cues in [20] and enhance the performance of even recent state-of-the-art appearance-based classifiers [39].

The difficulty for us lies in the fact that each photo is different in many ways. First, the number of people varies from photo to photo. This immediately prevents us to use a predetermined graphical model and node/edge potentials to classify all test photos. Secondly, human behaviors are quite complex. Any handcrafted rule is likely to have many exceptions, and the decision of when to or not to apply a rule is itself a difficult problem. Thirdly, it is not trivial to fuse the classification results from using facial positions and facial appearances, as they are inherently very different cues. In some scenarios, appearances are more useful than positions for gender/age classification whilst in other scenarios, positions are actually *more* useful. (See our Experiments Section.) In this work, we address each of these issues. Our method trains thousands of images within seconds, and enhances the performance of existing methods with high statistical significance.

Related work. Using non-appearance based contextual cues to aid in recognition tasks has started to receive attention in the computer vision community. In [77], pairwise social relationships (father-child, husband-wife, siblings, etc.)

were used to aid in face recognition tasks. In [23, 50, 71], photo co-occurrences of pairs of individuals were used, although the positions of the faces were not considered. In [2], facial positions were used, but only as a means of measuring the similarities of faces from different images. Further examples of using context cues include using first name priors for gender/age estimation [18], person matching across photos [69], and using captions for face identity [8].

Our method makes use of MRF to do gender/age estimation. Of course, using an MRF/CRF for classification tasks is not new. [23] modeled the faces of a photo as nodes in a CRF for face identity recognition. However, the potentials were derived from photo co-occurrences and friendship relations, and no facial positions were used. While their method is general enough to allow for any number of nodes (people), the specific selection in the work consisted only of 2-person photos. In [7], MRFs were used for various node labeling tasks. However, their work was primarily concerned with devising an improved MRF inference algorithm, and no classification accuracy improvement, if any, was reported. In a broader scope, many applications, such as foreground/background segmentation [10, 57, 75], optical flow [5], stereo imagery [63], multiview geometry [67], image denoise [6], among many others, can be viewed as variants of classification problems using MRFs. Solutions to these problems often appear in the context of energy minimization in which there is a smoothness constraint that neighboring nodes should be alike in some sense. However, this cannot be applied to facial gender and age classification, as the dependence of neighbors there works in a more subtle way than simple proximity. Indeed, close neighbors do not imply same gender!

From a didactic perspective, our work belongs to the same family as those that take advantage of structure to improve various learning tasks, commonly out-

side of human-related recognition. The use of contextual structure, even that from out-of-interest objects in the background, can aid in object recognition and scene understanding [45, 17]. Hoiem et al. [29] used 3D information to predict the locations of objects in an image. Even unlabeled objects or regions form structures that can be helpful [73, 13, 43]. Structure is essentially object relationship, and the benefit of jointly predicting related objects in an image cannot be overemphasized [28, 41, 46, 55, 17, 52, 54, 82, 9, 24, 59]. Our world is a highly structured one, and it makes sense to take advantage of it. For a survey of various contextual cues and their usage, see [16].

We use the dataset provided in [20]. The rest of the paper is organized as follows. In section 2, we describe our method. In section 3, we describe the experiments conducted to evaluate our method. In section 4, we demonstrate how well our method works in comparison to humans. Finally, we conclude in section 5.

2.2 Approach

We present our algorithm in this section. We model each photo as an MRF, where the nodes are the faces. We use a small validation set to decide, among the five canonical ones in Figure 2.10, which graphical model to use based on performance 2.7(b). Throughout this work, we will use Delaunay triangulation for gender classification and minimum spanning trees (MST) for age classification. For ease of illustration, we assume binary labels (gender), even though our algorithm works for arbitrary n-ary labels (e.g., age groups) in exactly the same way. Also, the number of faces in each photo does not have to be the same.

Training. The task is to learn the node potential functions Φ_F and Φ_M and the

edge potential functions $\Phi_{F,F}$, $\Phi_{F,M}$, $\Phi_{M,F}$, and $\Phi_{F,F}$. Different from what is widely done, we do not fit these potential functions as marginals in the MRFs, because doing so does not guarantee even a local optimum to the observed data likelihood from the MRFs.

First, consider a photo i from the training set as in Figure 2.2(e). Let the 2D positions of the faces be $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$. Following the baseline [20], the $\mathbf{x}_i$'s are mean removed (relative positions). Consider some labeling L over these faces. According to Figure 2.2(e), the unnormalized potential $T_i(L)$ is

$$\begin{aligned} T_i(L) \equiv\ & \Phi_{L(1)}(\mathbf{x}_1)\Phi_{L(2)}(\mathbf{x}_2)\Phi_{L(3)}(\mathbf{x}_3)\Phi_{L(4)}(\mathbf{x}_4)\times \\ & \Phi_{L(1),L(2)}(\mathbf{x}_2-\mathbf{x}_1)\Phi_{L(1),L(3)}(\mathbf{x}_3-\mathbf{x}_1)\times \\ & \Phi_{L(2),L(3)}(\mathbf{x}_3-\mathbf{x}_2)\Phi_{L(2),L(4)}(\mathbf{x}_4-\mathbf{x}_2)\times \\ & \Phi_{L(3),L(4)}(\mathbf{x}_4-\mathbf{x}_3), \end{aligned}$$

where $L(u)$ is the label of face u assigned by L. Let L_i be the ground truth labeling of photo i. The observed likelihood for photo i is then $p_i = T_i(L_i)/\sum_{L\in\pi_4} T_i(L)$, where π_4 is the set of all the possible $2^4 = 16$ labelings for a 4-person photo. The observed -log likelihood of all the training photos is

$$Q = -\log \prod_i p_i = \sum_i -\log p_i \tag{2.1}$$

The goal is to find the potential functions that minimize Q. Once we (A) choose the potential funtion models and (B) provide a formulation for the gradients of these models, we can apply gradient descent to 2.1 from an off-the-shelf software package [64] to find the (locally) optimal potential functions. We address (A) and (B) as follows.

Model. Let us denote by θ_l and $\theta_{l,m}$ $(l, m \in \{F, M\})$ the model parameters for the node and edge potential functions, respectively. For differentiable potential

functions, it can be derived that

$$\frac{\partial}{\partial \theta_l}(-\log p_i) = \sum_{u|u\in V_i} \frac{P[y_u = l] - I(L_i(u) = l)}{\Phi_l(\mathbf{x}_u)} \frac{\partial \Phi_l(\mathbf{x}_u)}{\partial \theta_l},$$

$$\frac{\partial}{\partial \theta_{l,m}}(-\log p_i) = \tag{2.2}$$

$$\sum_{u,v|(u,v)\in E_i} \frac{\partial \Phi_{l,m}(\mathbf{x_v} - \mathbf{x_u})}{\partial \theta_{l,m}} \times \tag{2.3}$$

$$\frac{P[y_u = l, y_v = m] - I(L_i(u) = l, L_i(v) = m)}{\Phi_{l,m}(\mathbf{x}_v - \mathbf{x}_u)}, \tag{2.4}$$

where $I(.)$ is the indicator function. Also, $P[.]$ denotes the MRF marginals, which can be computed from an off-the-shelf software package[65].

We choose the potentials to be Gaussian functions, each with parameters $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The gradients in (2.2)(2.4) can then be completed with

$$\frac{\partial \Phi_l(\mathbf{x_u})}{\partial \boldsymbol{\mu}} = \Phi_l(\mathbf{x_u})\boldsymbol{\Sigma}^{-1}(\mathbf{x_u} - \boldsymbol{\mu}) \tag{2.5}$$

$$\frac{\partial \Phi_l(\mathbf{x_u})}{\partial \boldsymbol{\Sigma}_{j,k}} = -\frac{1}{2}\Phi(\mathbf{x_u})\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}M^{j,k}\right) + \tag{2.6}$$

$$\frac{1}{2}\Phi_l(\mathbf{x_u})(\mathbf{x_u} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}\mathbf{M}^{j,k}\boldsymbol{\Sigma}^{-1}(\mathbf{x_u} - \boldsymbol{\mu}). \tag{2.7}$$

The case for the edge potentials is similar, with $\mathbf{x_u}$ replaced by $\mathbf{x_v} - \mathbf{x_u}$. Here, $\mathbf{M}^{j,k}$ is that square matrix $\mathbf{S}$ whose only nonvanishing entries are $\mathbf{S}_{j,k} = \mathbf{S}_{k,j} = 1$.

Node potential priors. Sometimes, we have a prior for some of the faces in the photos. These priors might come from another classifier, such as an appearance-based one. Our method allows for a way to combine these priors with the node potentials as follows. Let r_u^l denote the prior of face u for label l. We then modify the node potential as $\Phi'_l(\mathbf{x}_u) = (1-\lambda)r_u^l + \lambda\Phi_l(\mathbf{x}_u)$, where $0 \le \lambda \le 1$. λ can be tuned via cross validation. Throughout our experiment, we fix $\lambda = 0.9$. In addition, we set all the initial Gaussians idential, so that at iteration 0 of the gradient descent our method behaves the same as the priors.

Test. Given a test photo, we use the potential functions to compute the marginal probabilities of each face using off-the-shelf MRF inference algorithms, and take the maximum one as the predicted label.

Learning. It is worthwhile to point out the performance difference between our learned potentials and fitting the potential functions as marginals, which is commonly done [62, 79, 72, 20, 77, 7]. In Figure 2.2, (a) - (d) are the age classification results over different methods, with 95% confidence interval (CI) of the performance improvement provided. (See the caption.) (c) $\rightarrow$ (d) gives the performance improvement CI of our method over marginal fitting of potentials. This is consistent with [66], which recently pointed out the deficiencies of MAP estimates in certain contexts [44, 51, 53].

In the figure, (a) $\rightarrow$ (b) and (c) $\rightarrow$ (d) show the improvement from learned over marginal fit potentials, while (a) $\rightarrow$ (c) and (b) $\rightarrow$ (d) show the improvement from using the MRF structures. In the case of not using appearance-based priors, (a) is exactly the baseline [20]. (d) is our method, which is therefore a strict generalization of [20]. Interestingly, the CI lower bound of either (b) or (c) alone is $< 1\%$, while that of our method is $> 4.5\%$. When not using any appearance-based priors, we set the initialization of gradient descent to the marginal fit functions.

Abbreviations. We denote PP as our ProPosed method, and BL as the BaseLine method [20], both of which only use facial positions without any appearance-based priors. We denote APR as using APpeaRance-based features (only) for classification based on [39]. Finally, we denote PP.APR and BL.APR as combining the appearance-based priors with the corresponding position-based algorithm.

Difference (%)	95% Conf. Int.
PP - BL	(+5.47%, +6.13%)
PP.APR - APR	(+2.53%, +2.94%)
PP.APR - BL.APR	(+1.30%, +1.69%)

Table 2.1: Performance improvement of our methods in gender classification.

2.3 Experiments

In this section we describe the experiments performed to evaluate our method. Following [20], the performance metric is accuracy. I.e., the number of faces that are correctly predicted. All of the performances reported are averages over at least 20 independendent splits of data into training and test sets, with 95% CI's of the performance difference using paired two-sample t-tests provided to demonstrate that the improvement from the proposed algorithm is statistically significant. Unless otherwise noted, in each training/test split, all the classifiers (baseline and proposed) are trained afresh using the corresponding training set. Figure 2.11 shows some results of PP. They are picked from the instances in which our algorithm works relatively well, to show what the relative facial positions of typical useful cases look like.

2.3.1 Gender Classification

In this task, we aim to predict the gender of each face from the set of test photos. **Using appearance.** We carry out 20 random splits of 3522 photos into 60% training and 40% test sets. See Figure 2.7 for the distribution of the number of faces per photo. Since we are not exploring face detection algorithms in this work,

we directly use the position and size of each face as inputs as provided from the ground truth. In each training/test split, we compare our method with the baseline algorithms APR and BL.APR. Figure 2.3(b) summarizes the average performance over 20 runs. Table 2.1 shows the statistical significance of our improvement.

No appearance. Using facial position cues only, without any appearance-based features, gives a prediction result that is surprisingly better than a random classifier [20]. Here we compare our PP with BL without using any appearance features. Figure 2.3(a) and Table 2.1 summarize the results.

Training size effect. It is interesting to investigate the number of training photos it would take to achieve the 65% accuracy in Figure 2.3(a). To estimate this, we randomly select 40% of the 3522 photos as the test set, and use k of the remaining 60% as the training set, where k runs from 15 to $2113 = 3522 \times 60\%$ at 10 evenly spaced samples (Figure 2.4(a)). For each value of k, we perform 20 random training/test splits and average out the accuracies. From the figure, we see that it takes ≤ 250 training photos for PP to achieve the 65% accuracy. This suggests that, for our method, the benefit of position cues in gender classification does not require as many training photos as using appearances. Indeed, when we artificially restrict the training set size for our method to be no more than 250 (whilst still maintaining the training set size for APR as the x-axis of Figure 2.4(a)), we obtain a curve that is almost identical to the curve PP.APR in the figure. In other words, the improvment of our method over an appearance-based algorithm requires almost an order of magnitude less training set than the appearance-based algorithm.

It is worthwhile to point out that, when there are much fewer (< 150) training photos available, our method performs significantly *better* than APR. Figure

2.4(b) shows the same experiment setting of 20 random splits, where k now ranges from 15 to 100. There, we see that PP and BL perform significantly better than APR. That is, using only positions significantly outperforms using appearance features. Evidently, this improvement comes from modeling positional neighbors' gender dependency, and demonstrates that the gender of a person and that of his/her neighbor are not independent events.

Improvement factors. We are interested in figuring out the circumstances in which our method performs better than the baselines. The two factors we consider are the number of people present in a test photo and the sizes of the faces. Once again, the results are all based on 50 random splits of 60% training set and 40% test set.

Figure 2.4(c) shows the performance *differences* of our method to the baselines APR and BL.APR with respect to the number of people in a photo. As we increase the number from 4 to 8, the improvement of our method over the baselines increases. This is expected, as the presence of people naturally implies richer gender dependence among the neighbors. Interestingly, the performance improvement decreases as we further increase the group size. A possible reason is that there are much fewer photos with ≥ 10 people in them (Figure 2.7(a)), and a smaller number of training set for these photos makes it less sufficient to learn neighbors' gender dependency therein. Note, however, that the difference is still > 0. That is, our method still outperforms the baselines in these cases.

Figure 2.4(d) shows the performance improvement of our method with respect to face sizes. Here, we model the face size as the eye distance in absolute number of pixels. The improvement over APR is higher for smaller faces. This is expected, as faces with a lower resolution would provide less information as an appearance-based cue, but not as a position-based cue. Interesetingly, we ob-

serve a similar trend for the improvement over BL.APR.

2.3.2 Age Classification

We partition age into 4 groups (child, youth, adult, senior) and formulate the task as a 4-way multinomial classification problem. This is slightly different from [39], where binary classification result of each attribute was individually reported. We select a subset of 712 photos from the 5080 photos. The subset is chosen so that the age distribution of the faces is approximately even across the 4 age groups. We use linear programming to ensure that no subset larger than 712 photos constitutes an even age distribution (selecting at photo level). We carry out 20 random splits of these photos into 80% for training and 20% for test. The result is summarized in Figure 2.5 and Table 2.2. Our method outperforms the baseline either with or without using an appearance-based age classifier as a prior. The 95% confidence intervals from Table 2.2 conclude that the difference is statistically significant.

Figure 2.8(a) shows the effect of training size to age classification accuracy of various algorithms. The setup is the same as in gender classification, and each data point is averaged over 20 random training/test splits. Here, we again observe that APR does in fact perform far worse than PP and BL, which do not use any appearance features. Also, with a small number (< 600) of training photos, our method is the only one to achieve an accuracy that is more than twice that of a random classifier. Unlike the case of Figure 2.4(a), it is not obvious whether we can improve the performance of PP with more training photos.

The accuracy of 34% in Figure 2.8(a) of APR is lower than the 66% reported in

Difference (%)	95% Conf. Int.
PP - BL	(+8.50%, +11.76%)
PP.APR - APR	(+3.23%, +3.92%)
PP.APR - BL.APR	(+2.24%, +2.90%)

Table 2.2: Performance improvement of our methods in age classification.

Figure 2.5. In the setting of Figure 2.5, APR is trained using a larger set of both the 80% of 712 photos *and* the faces from the remaining 5080 - 712 = 4368 photos whilst maintaining even age distribution. (Although PP and BL are genuinely trained using only the $712 \times 80\%$ photos.)

2.3.3 User Feedback

A fundamental difference between the baseline and our method is that the former does not make predictions that depend on the posterior likelihoods of the other faces in the same photo, while the latter does. For our method, the provision of the ground truth labels of some faces in a photo affects the classification results of the other faces in the photo. In this experiment, we demonstrate how our algorithm performs when the labels of some faces in a photo are given. Specifically, for each test photo i, where $|V_i| \geq 6$, we randomly partition V_i into two sets K_i and U_i, where $|K_i| = 5$ and $|U_i| = |V_i| - |K_i|$. We test the classificaton performance over the faces in U_i using PP (without any appearance cues) by divulging the true labels of m (number of feedbacks) randomly chosen faces in K_i, where $0 \leq m \leq 5$. Figure 2.6 shows the results of this experiment for both gender and age classification. These results give a clear evidence as to the mutual dependency of the labels of the faces in a photo. By harnessing such dependency appropriately, classification performance can be improved. Note in Figure 2.6(b)

that revealing the age of just one person in a group of ≥ 6 people is sufficient to boost the age classification performance by 15%. This feedback mechanism is amenable to interactive applications where the human user partially provides some of the face labels in a photo. The age classification for the user feedback experiment is done over a finer 7 age groups (0-2, 3-7, 8-12, 13-19, 20-36, 37-65, 66+), which make classification harder.

2.4 Human Studies

In order to compare the performance of our algorithm with humans, we design an experiment to test humans' ability to classify gender/age based on limited information. The experiment is conducted as follows: For photos with more than three people, we show the gender/age of three of the people. For photos with two or three people, we only provide the gender/age of one person. The human respondent is required to interpret the gender/age of the rest of people in the photo based on the position and size of the faces as shown in Figure 2.9. As the respondent chooses the gender/age of a face in the synthetic photo as in Figure 2.9(b), we display his/her past choices on the screen so that they may be used for guessing the label of the current face if the respondent wishes to do so. The respondent cannot go back to change his/her past choices.

Figure 2.8 shows the accuracies of the proposed algorithm and humans. The proposed algorithm outperforms humans in gender classification and gives competitive results for age classification. This result shows that, without appearances, our algorithm is at least as effective as humans.

The main difference between humans and our algorithm is that humans tend to interpret the social relation of the people in the photo before they make the pre-

diction on gender/age. If the given faces are all in the same age range, humans tend to guess the rest of people in the same photo as in the that age range. However, if the ages vary within the photo, people tend to assume a family photo and make the predictions based on this prior. This tendency explains why human performs slight better than the proposed algorithm on age tasks, since our algorithm does not separate the age distribution from family photos and group photos, which can be quite different. Despite the advantage of various social and cultural knowledge priors, however, the human respondents do not perform significantly better (age), and in fact performs worse (gender), than our algorithm.

2.5 Conclusion

In this paper, we model gender and age dependence of people in a group photo using graphical models, and show that it significantly outperforms previous works. We also provide a parameter learning methodology which, different from what is usually done in graphical model applications, does yield the potential functions that maximize the observed data likelihood. In various experimental settings, our method performs significantly better than current methods. What is more, our method requires no parameters to tune (as they are automatically learned), trains thousands of photos within seconds, and can outperform humans.

Of course, there are limitations to our work. For example, it'd be worthwhile to investigate into choosing different graphical structures at the photo level. Also, the choices of the potential functions may be extended as well. We believe that the full benefits of social-based proximity are yet to be tapped.

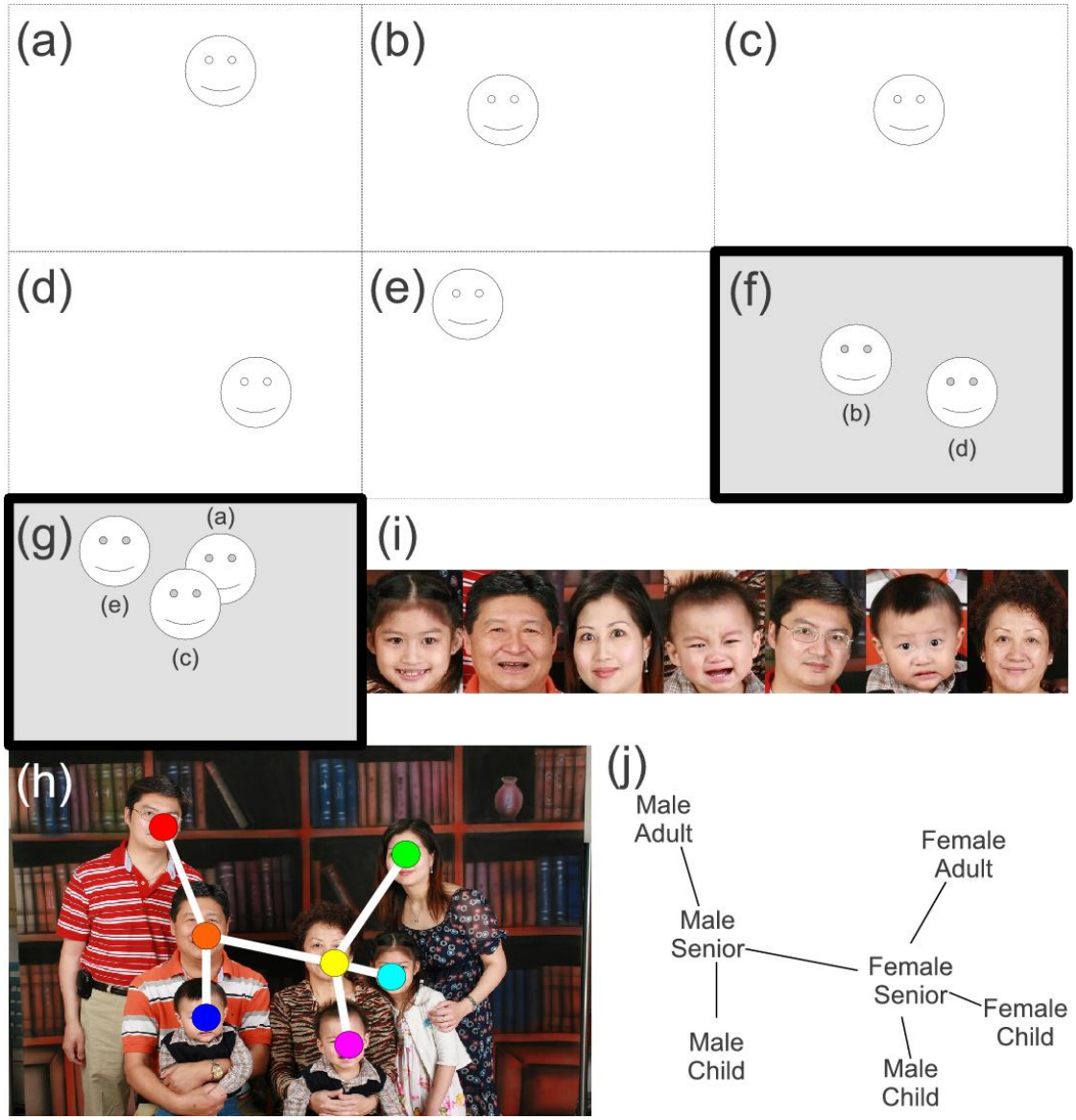


Figure 2.1: Without appearances, position cues for estimating gender and age are weak (a) - (e), difficult even for humans. With the extra information that (b) and (d) belong to a two-person photo (f), and that (a), (c), and (e) belong to a three-person photo (g), it becomes easier to estimate that (b) is male, (d) is female, (e) is an adult male, (a) is an adult female, and (c) is a teenager. In this paper, we harness this extra information (h). Incorporating the appearance-based cues (i) in a principled way, we demonstrate that our method can (j) further enhance current appearance-based state-of-the-art.

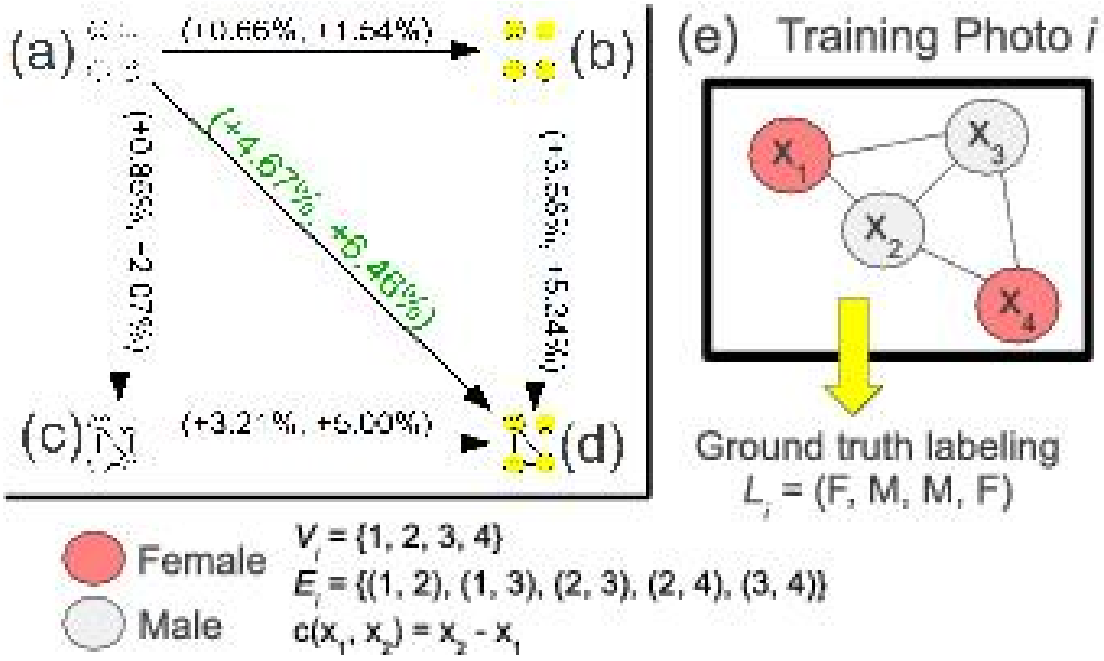


Figure 2.2: (a) and (c) use marginal fit potentials, while (b) and (d) use learned potentials. (a) and (b) use no MRF edges, while (c) and (d) use MRF edges. (e) illustrates the idea of our algorithm.

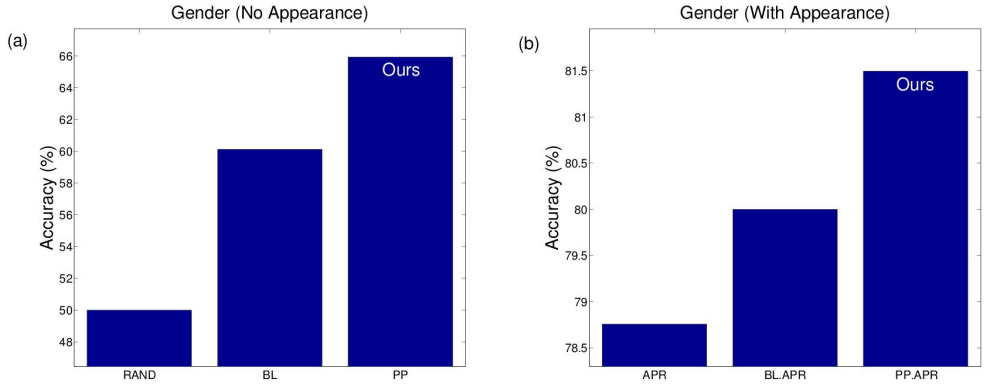


Figure 2.3: Gender classification result averaged over 20 independent runs.

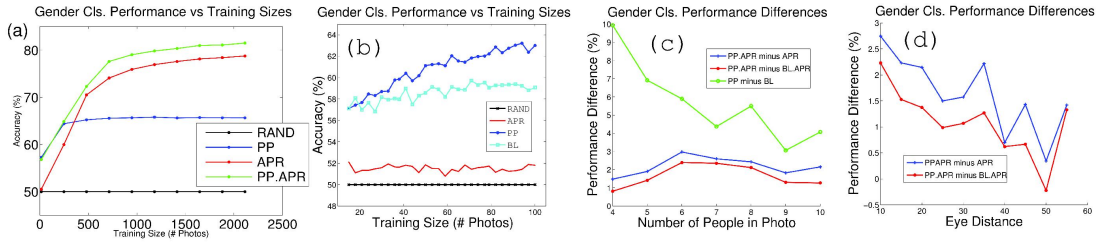


Figure 2.4: Except RAND, each data point in (a) - (d) is averaged over 20 independent runs.

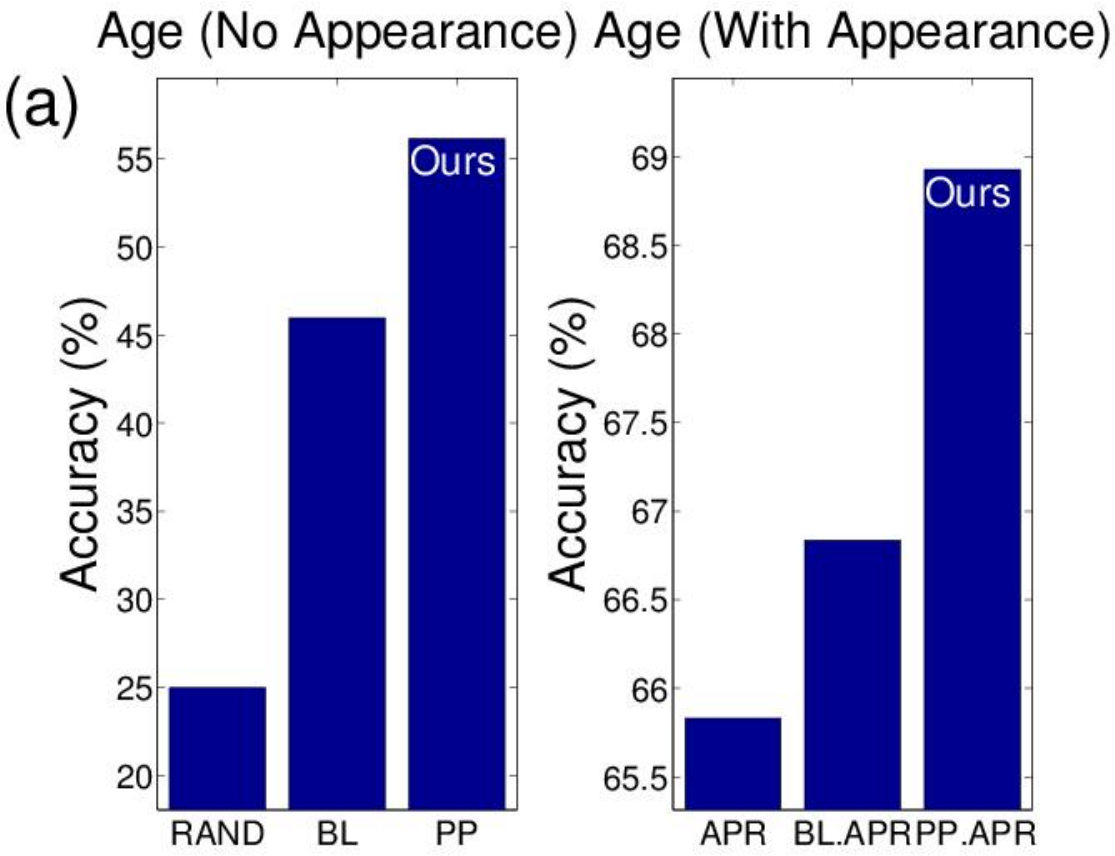


Figure 2.5: Age classification result averaged over 20 independent runs.

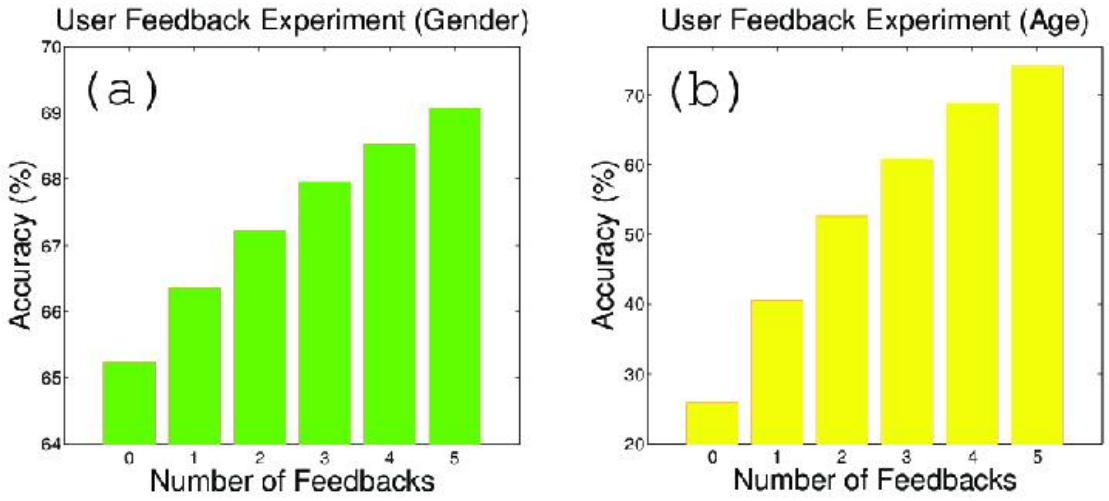
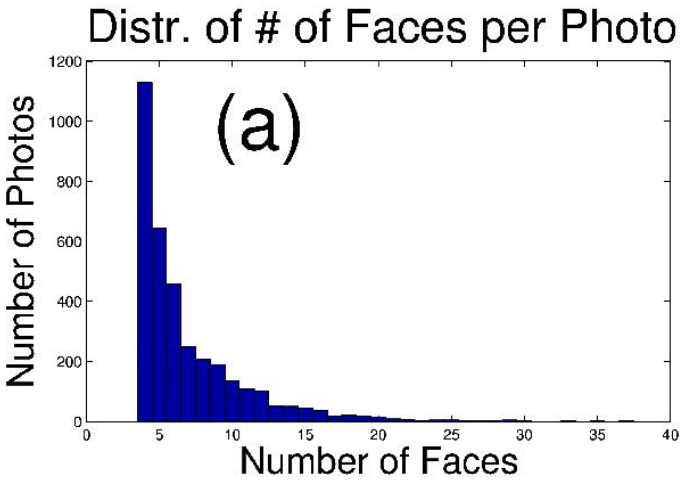


Figure 2.6: User feedback experiment. Each bar averaged over 50 independent runs.



(b)

Graph	Mean Accuracy
(1) No edges	61.2%
(3) Complete graph	61.7%
(4) Star graph	63.9%
(5) Short Ham. Path	65.0%
(2) MST	65.3%
(6) Delaunay	65.8%

Figure 2.7: (a) Distribution of the number of faces. (b) Gender accuracy with different graphs, averaged over 50 independent runs.

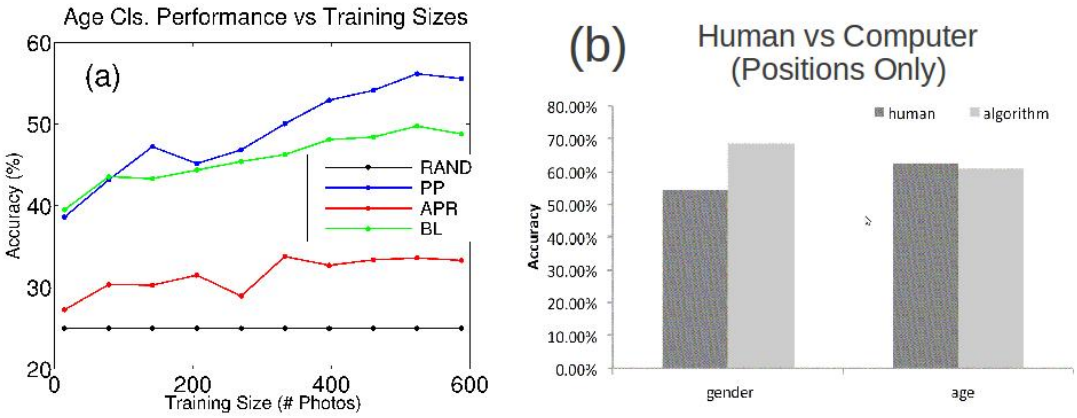


Figure 2.8: (a) Age cls. performance over training sizes, averaged over 20 independent runs. (b) The comparison between human performance and the proposed algorithm. Results averaged over 10 human respondents for age, and 13 for gender.

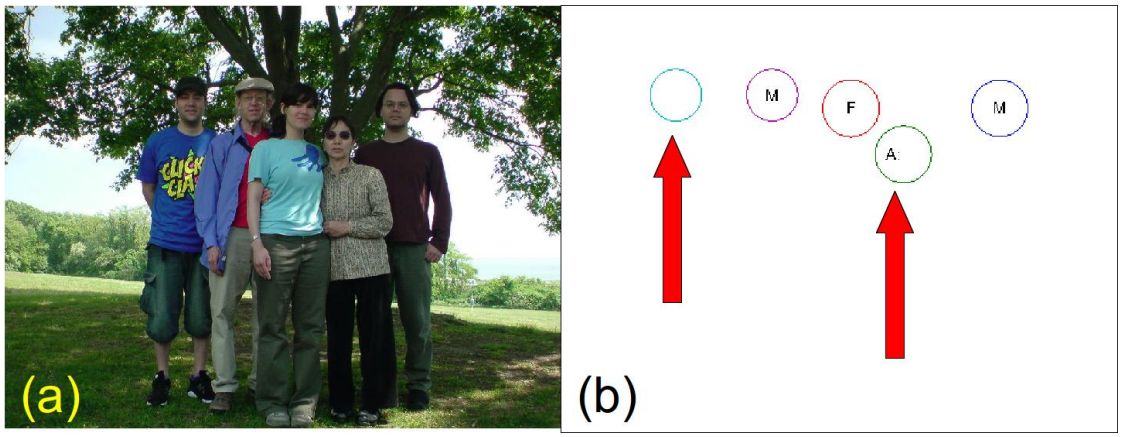


Figure 2.9: Fig. (a) shows the original photo. Fig. (b) shows the picture we provide to the respondent for interpreting the gender.

Figure 2.10: Graphical model structures including the (a) complete, (b) star, (c) shortest Hamiltonian path, (d) minimum spanning tree, and (e) Delaunay triangulation graphs. (Better viewed in color)

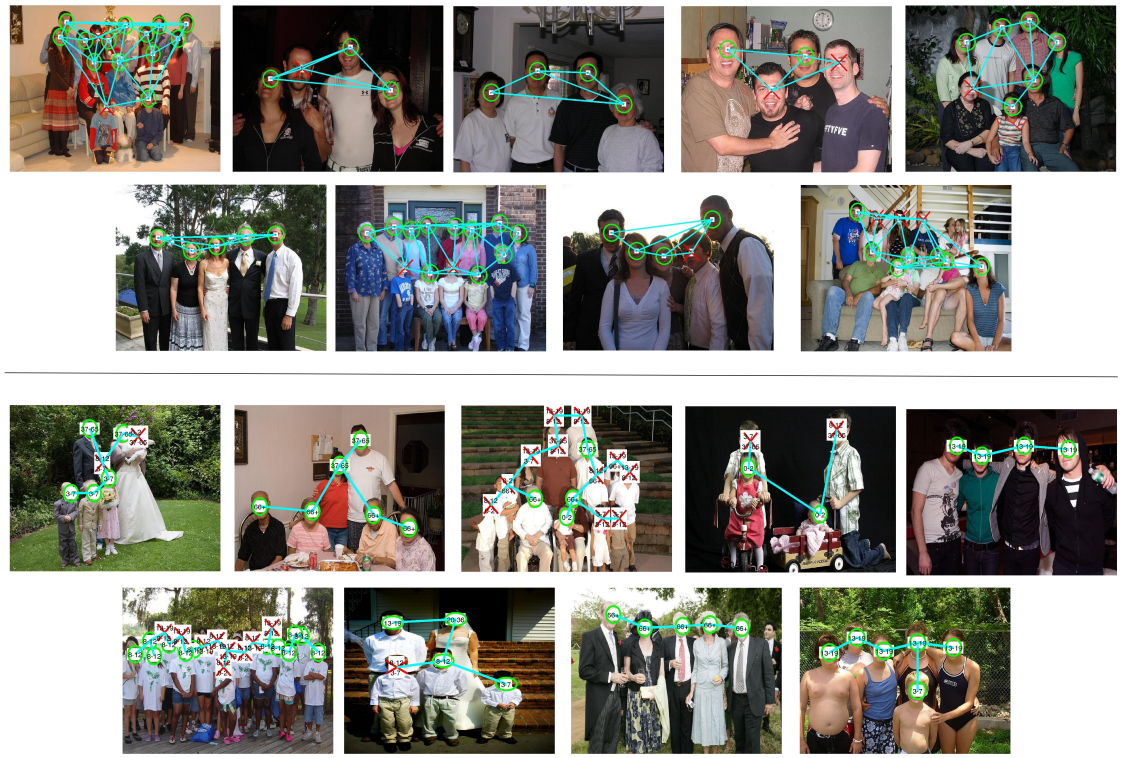


Figure 2.11: Results of gender (top) and age (bottom) classification using positions only (without any appearance features). The MRF structures from PP are shown. Circles are correct predictions and crosses are wrong predictions. For age, the top number is the ground truth, and the bottom one is the prediction (shown only if different from the ground truth), with the finer 7 age groups. Best viewed in electronic version.

CHAPTER 3

RELATIVE DEPTH ESTIMATION IN A GROUP PHOTO

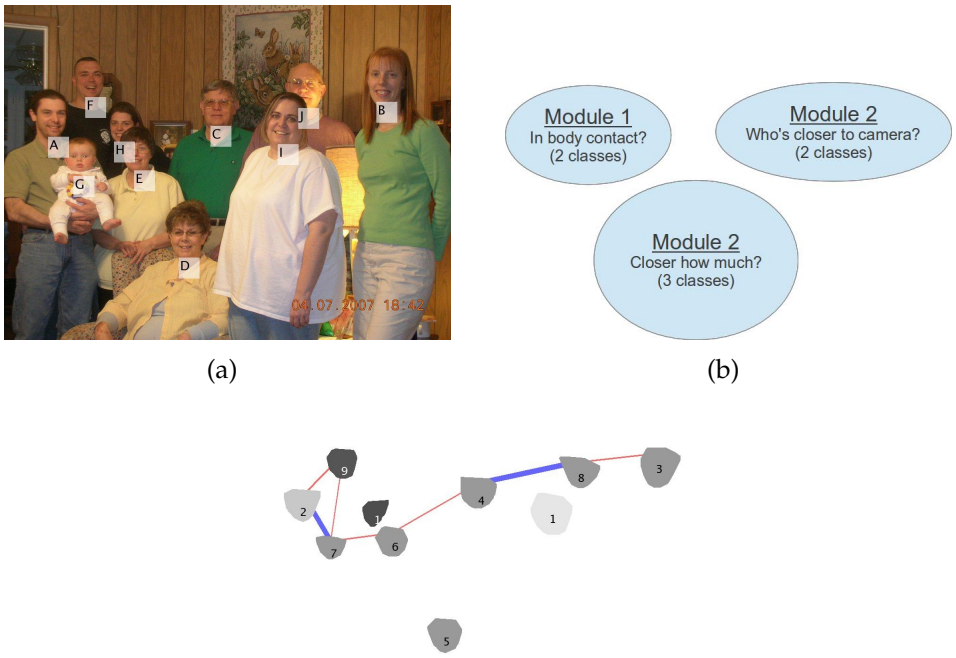
3.1 Introduction

The "group shot" is a photograph of a group of people and group shot instances are captured millions of times each year. When a human looks at a group shot, she easily gains a sense of the members of the group and the space that they occupy. Not only can she locate the persons within the image, but also understands their positions in space and their interactions with one another. In short, the 2D image of a group of people is interpreted as a 3D arrangement of people. This captures the goal of this paper: to interpret group images in 3D space by inferring the 3D position (i.e., to estimate z) of each person in the image, and whether people are in physical contact.

As a concrete example, consider Figure 1. The persons on the right of the image (B and I) are the closest to the camera plane, while person F is farthest from the camera. No simple heuristic, examining features such as the face y-coordinate or the face size, can by themselves deduce this arrangement.

By interpreting the geometry of a group image, we are poised for applications that benefit from understanding human positions and interactions, not just in two dimensions, but in three. Past work [21] has shown that social factors influence even 2D positional arrangments of persons within a group image, and we expect that this work will lead to improved methods for exploiting social context in groups.

On one hand, this problem can be viewed as a special case of inferring depth from 2D images [61, 31, 35] that has inferred a great deal of attention in recent



(a) (b)

(c)

Figure 3.1: Given an input image (a) our algorithm estimates the z-coordinate of each person, rendered in (b). We model this problem as a joint classification of all the 3 modules. (c) The result is the prediction of the relative distance of each person to the camera (encoded in grayscale) and whether pairs of people are in physical contact (blue edges). The numbers indicate the sorting of the people in their z-coordinates.

years. We rely on the insights of those general depth-from-image works for inferring the depths of persons in our group images. Instead of inferring depth based on texture cues, scene priors, and object labels, we infer the depth from human-relevant features related to face sizes, facial landmarks, (2D) positions. In addition, we detect person occlusions via shoulder-torso HOG features encompassing neighboring people as strong cues of the ordering of adjacent people.

We propose a model that combines several types of features to infer: depth orderings of sub-groups of people in the image, relative depth (i.e., Δz) between

pairs of people, and the presense physical contact between persons in the image. To encourage global consistency, we classify all of these tasks jointly. Finally, we use linear programming to assign a z-coordinate to each person.

In summary, the main contributions in this work are:

1. We propose the problem of jointly finding relative depth estimates and physical contact for group images of humans, and propose novel HOG-based features for detecting human occlusions.
2. We unify depth order, distance, and physical contact modules for z-coordinate estimation, which guarantees global consistency.
3. A dataset containing relative z-coordinate distance, and body contact dataset of over 5000 consumer group photos.

3.2 Related Work

The goal of our work is to produce a scene interpretation of an image of a group of people, including each person's depth from the camera plane, and a prediction of which pairs of people are in physical contact. In this regard, our work is related to previous works that perform inference on attributes or quantities regarding pairs or groups of people, and works that perform 2D to 3D conversion.

Groups of People: Our work is related to the recent works from Wang and Ai [78], Jia et al [35], Gallagher and Chen [19], and Yang et al [81].

In [78], the goal is to perform clothing segmentation in a group photo. This work has overlap with ours in that they find instances of pairs of people where

one occludes (blocks) another, but they do not find a relative depth estimating for each person in the scene. Because their goal is segmentation (and not 3D interpretation), their algorithm makes no inference regarding pairs of people who do not have touching segments. In contrast, our method finds a depth ordering and relative depth location for all persons in the group image, and infers the position even of persons who neither occlude, or are occluded by, another person (e.g., person *B* in Fig. 4.1(a).) The work of [19] groups people in a photo into rows but does not provide a direct estimate of the the relative depth between persons, or provide a depth ordering of people (or rows of people). In [81], the goal is to classify touch codes given pairs of people in contact, such as whether two people are holding hands. However, in a group photo, it is a common occurrance that many people are mostly occluded (as also observed by [78]), and inferring whether a pair of people are touching, by directly observing image content that may be occluded, proves to be difficult. We jointly infer when people are in physical contact, and the relative distance (in z).

Depth from 2D: In the most general sense, our goal of finding the depth ordering and relative depth of people in a group image shares can be seen as a special case of the general depth-from-2D problem that has received a great deal of attention in the literature [60, 30, 35]. These works often find superpixels, then carry out depth inference based. Examples include Hoiem et al [31] and Liu et al [47], which use semantic labels such as "ground", "sky", etc. to estimate depth from context. Saxena et al [61] learns a regression for depth based on superpixels. These works focus on interpreting the 3D world behind the 2D image in the general case, but make assumputions that objects are grounded, planar, and front-facing. There are instances of 3D interpretation that exploit the special

(a) (b)

Figure 3.2: (a) Sample Mechanical Turk HIT interface. (b) These distances are inconsistent with $A <_z B <_z C$. Our model addresses this problem by using a z-coordinate assignment algorithm that guarantees global consistency.

structure of the scene: for example, in [58], object classes, prior knowledge of object size, and outlines are used for converting a scene to a 3D model. In [27], the structure of indoor rooms is inferred by exploiting the structure of rooms and representing objects (e.g., beds) as boxes. Our paper's goal is to exploit the special structure of our scenario: that people stand in predictable arrangements with respect to one another, and by learning this, we can infer the relative depth z of people in the group image.

Indeed, depth estimation from a single monocular image has been well studied so far.

3.3 Data Collection

We use the dataset of [21], which consists of 5080 consumer photos with the ground truth face positions provided. For each photo, our goal is to classify pairs of people in the photo by whether one is closer to the camera than the other (2 classes), their absolute distance in the z-axis (3 classes for *less than 8cm*,

between 8cm and 100cm, and *greater than 100cm*), and whether they are in body contact (2 classes). We pay human workers from Amazon Mechanical Turk to provide answers to these questions as the ground truth. To be more robust against human errors, we only enlist Mechanical Turk Master workers. In addition, each question receives answers from several different Master workers, and the majority vote is taken. Fig. 3.2(a) shows a sample HIT interface.

A pair of people is selected as a Turk question candidate only if they form an edge in the Delaunay triangulation of the faces in the photo, or if one person is one of the three closest neighbors of the other person. Intuitively, this restriction helps keep the number of questions down while ensuring that nearby faces have ground truth data collected.

Even with the measures above, it is still possible that a photo contains self-inconsistent ground truth. A simple example is shown in Fig. 3.2(b). To address this issue, we apply the z-coordinate assignment algorithm (see the Algorithms section) to the data collected from Mechanical Turk, and derive the final pairwise ground truth data from these z-coordinates.

3.4 Body Contact and z-Coordinate

It is perhaps not surprising that the z-coordinates of a pair of people in a photo are not independent to whether they are in body contact. After all, the distance (in 3D space) between the people can affect whether they are in body contact. To investigate this dependency in more detail, we collect pairs of people in a photo that are closest neighbors of each other in the pixel plane. For each such pair, we crop out a region of the photo containing them. For simplicity, we assume that the person on the left has a lower y- coordinate. Whenever this does not hold,

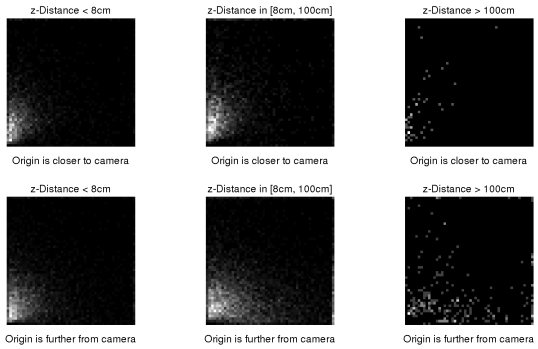
we horizontally flip the cropped region. Examples are shown in Fig. 3.4.

From Fig. 3.4, it is clear that classifying body contact is a hard problem. For example, in Fig. 4, both (a) and (b) have similar appearance, yet one pair of people are in contact while the other are not. In addition, there are very different ways for which two people may be in contact, such as by hands, entire (h), etc. (k) and (l) show pairs of people under severe occlusion and clutter, which make the classification of body contact very difficult. The state-of-the-art work [81] classify each pair of people by the mode of touch they are in. Here, we are interested in classify simply whether two people are in touch (contact). In particular, we focus on how knowing the positions in 3D of each person in the photo can help this task.

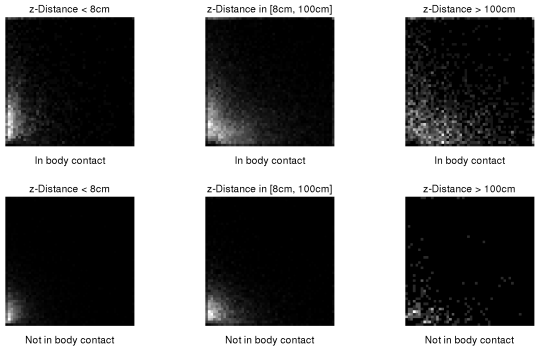
Fig. 3 supports the fact that the 3 modules are dependent with one another. In (a), we plot the distribution of positive body contact based on the ground truth of the other labels (depth order and relative distance), for a total of $2 \times 3 = 6$ distributions. Clearly, we see that the distribution of body contact depends on depth order and relative distance. Similar results are shown in Fig. 3(b), for the distribution of depth order based on body contact and z-Distance.

3.5 Algorithms

Within each image, the pairs of people selected are those that are either closest neighbors or correspond to an edge in the Delaunay triangulation of the faces in the pixel space. For each pair, we extract several types of features. Then, we train a single classifier that jonitly learns all of the body contact, depth order, and relative distance relations of each pair. Finally, from the predicted pairwise depth order and relative distances, we estimate the z-coordinate of each face.



(a)



(b)

Figure 3.3:

(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l)

Figure 3.4:

3.5.1 Feature Extraction

There are two types of features that we use in this work: the appearance-based features and contextual features.

Contextual Features

For each image, we take the ground truth of the face bounding boxes and apply [83] to find the facial landmarks for each face. It also detects the pose angle of each face. As in Section 4., all of the faces on the left are made so that it has a lower y-coordinate in the pixel plane. For each pair of people i and j to consider, we extract several non-appearance contextual features as follows.

Areas. From [83] the facial landmarks of each face essentially form a polygon. We estimate the area A_i of each face i by that of the convex hull of all the landmark points of the face. Finally, we take the ratio of these areas as one feature: A_i/A_j.

Coordinates. We use the xy-coordinates x_i and y_i of each face i in several ways. Firstly, we use the normalized differences $(x_j - x_i)/(A_i + A_j)$ and $(y_j - y_i)/(A_i + A_j)$ as features. Also, we use the coordinates normalized by the image width W and height H as additional features: $(x_i/W, y_i/H, x_j/W, y_j/H)$. Finally, we also use the distances $|x_j - x_i|$, $|y_j - y_i|$ in three forms: unnormalized, normalized by image dimensions, and normalized by face areas.

Pose Angle. We take the pose angle produced from [83] of each face binned at 90° intervals as another feature.

Appearance features

The appearance features we use are the HOG features, using the implementation [74]. For each pair of people, we first crop out the region in which their face positions are fixed. Then, we extract the HOG features from each of these cropped regions.

3.5.2 Classification

Finally, we consider this as a $2 \times 2 \times 3 = 12$-way classification problem (2 labels for body contact and depth order each, and 3 labels for relative distance). Each of the 12 labels encode for a particular configuration for each module. There are two primary reasons that we choose to predict the labels jointly instead of training one separate classifier for each module. Firstly, as discussed in Section 4., these modules are not independent. Training them jointly in this high-order manner will help capture this dependency. Secondly, we have ample training data, and a 12-split of them does not incur the problem of vanishing empirical data occurrence.

We use SVM to do the classification, with RBF kernel with the parameter γ set to 1 throughout and perform a 10-fold cross validation on C.

3.5.3 z-Coordinate Assignment

At the second stage, suppose that E is the set of pairs of people appearing in some set, with the notational convention for $(i, j) \in E$ to mean that person i is closer to the camera than person j. Once the depth order and relative distance

Depth Order	**Closer**	**Further**	
Precision	79.24%	78.57%	
Recall	78.89%	78.93%	
Overall Accuracy			78.91%

Table 3.1: Pairwise performance results for depth order.

classes of (i, j) are known, the computation of the z-coordinate of each peron can be formulated as the following linear program over nonnegative variables.

$$\min_{z,\xi,\tau} \sum_{(i,j)\in E} \xi_{(i,j)} + C\tau_{(i,j)} \tag{3.1}$$

$$\text{where } \forall (i, j) \in E : \tag{3.2}$$

$$|z_i - z_j| < ub + \xi_{(i,j)} \tag{3.3}$$

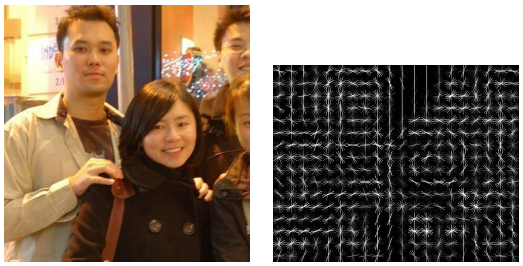
$$|z_i - z_j| > lb - \xi_{(i,j)} \tag{3.4}$$

$$z_i - z_j < \tau_{(i,j)} \tag{3.5}$$

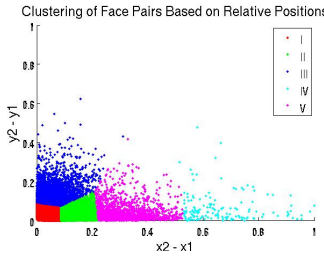
Here, Ineq. (3.3)(3.4) encode distance upper and lower bounds derived from the relative distance class. Likewise, Ineq. (3.5) encodes the depth order class. ξ and τ are slack variables. Clearly, this linear program has an optimal solution of value 0 iff all the pairwise classes are globally consistent. Otherwise, the linear program would still find the z-coordinates that violate the classes as little as possible through non-vanishing slack variables. C is a positive constant that controls the preference of satisfying depth order over relative distance. In our experiments, we set $C = 1$.

3.6 Experiments

In our experiments, we randomly split the data set of 5080 photos from [21] into 60% training set and 40% test set. These amount to a total of 32072 training pairs



(a) (b) (c)



(d)

Figure 3.5: (a) The original cropped image of a pair of people. (b) The HOG rendering. (c) The sparsity of SVM makes many of the feature points vanish. (d) 5 clusters of the normalized *xy* positions of pairs of people in our data set.

Body Contact	**No Contact**	**In Contact**	
Precision	72.68%	63.35%	
Recall	85.00%	44.80%	
Overall Accuracy			70.26%

Table 3.2: Pairwise performance results for body contact.

Relative Distance	**(1)**	**(2)**	**(3)**	
Recall	50.64%	61.31%	27.03	
Precision	47.50%	75.62%	0.0065	
Overall Accuracy				57.83%

Table 3.3: Pairwise performance results for relative distance. (1) is less than 8cm, (2) is between 8cm and 100cm, and (3) is greater than 100cm.

and 13914 test pairs. For each of these pairs, we further divide them into 5 group by their normalized xy-coordinates. We use the k-means clustering algorithm to obtain the group assignment. Fig. 3.5(d) shows a typical clustering result. By dividing the pairwise data into different groups, each classifier can be taylored to its specific subset of data. For each group, we resize each cropped image to the mean crop size in the group before extracting the HOG features.
Tables 1-3 summarize the pairwise performance results of the three classification modules.

3.7 Conclusion

In this work, we have proposed a novel problem of estimating the z-coordinate of each person in a group photo. In addition, as part of the related task we incorporate the classification of body contact into the problem, as it is highly related to the positions of people in 3D. Our results have quantitatively confirmed the solvability of this novel problem, and establish as a baseline to which future methods can be compared against.

CHAPTER 4

FACE-GRAPH MATCHING FOR CLASSIFYING GROUPS OF PEOPLE

4.1 Introduction

People often gather for a photo shot for an underlying social reason. Past works have shown that the spatial arrangement of the faces in a photo provides useful cues as to predicting certain attributes of the faces ([3, 22]). In addition, when harnessed properly, the pairwise spatial positions of faces can also give useful information in predicting the relationships of the individuals in the photo ([22]). Of course, the social relationships and events under which a photo was taken can affect how we humans might categorize the group. Motivated by this observation, our goal in this work is to investigate the relationship between the spatial arrangement of faces in a photo and the type of group that has assembled.

Consider the four photos (a) - (d) in Fig. 4.1, in which all visual content is removed except the faces, their relative sizes, and some age or gender clues. Using only this information, the reader can attempt to categorize each photo with an appropriate label on the right. Compare your answer to the results shown in Fig. 4.2. How many photos did you classify correctly?

The reader will probably score much better than the 25% random choice. As these simple examples show, facial arrangement and attributes provide an important cue useful for photo type classification. In this work, we demonstrate the usefulness of this cue for group photo classification.

Related work. The usefulness of facial arrangements have been explored before for predicting attributes on single people [22] and relationships between pairs

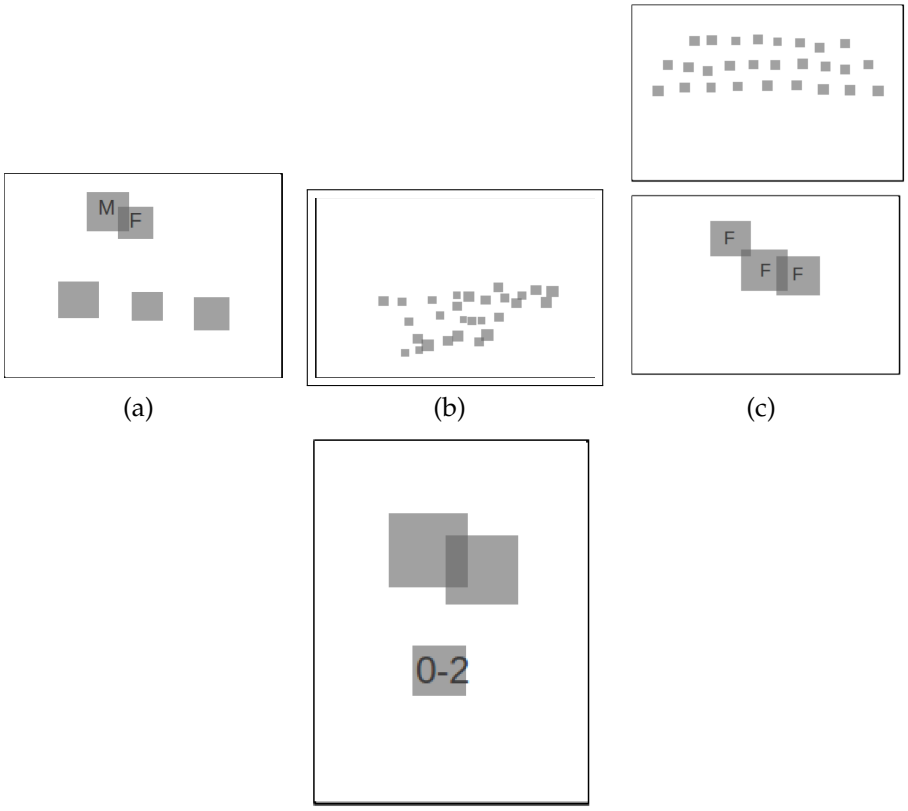


Figure 4.1: The label space is (1) Family, (2) Group Field Trip, (3) Sports Team, and (4) Friends Hanging Out

of people [68, 80, 76]. Also, in [3], facial arrangement was used to measure the similarity of two photos. Although the task was for image clustering rather than classification, and the goal was targeted towards human-subjective ranked retrieval assessment, the motivation that facial arrangement has to do with photo similarity is the same. [22] used the least square fit of face sizes and positions to detect group dining photos. Most of these works involve the use of facial positions and attributes. More traditionally, image classification is often conducted with appearance-based features. Examples include face attribute classification [38, 40], occupation prediction based on the kind of clothing a person wears [70], cultural type and urban tribe classification [49]. However, we believe that

Figure 4.2: Answers to Fig. 1. The label space is (1) Family, (2) Group Field Trip, (3) Sports Team, and (4) Friends Hanging Out

classifying consumer photos should be based on the humans. After all, they are the protagonists of a story the photo tries to convey.

4.2 Ground Truth and Data Collection

Previous work provides some datasets of photos in which something about the photo type is known. For example, [22] gives a rough photo categorization of group, family, and wedding. Often in the previous work, the photo type was derived from the tags that were associated with each photo or the search query terms used in retrieving the photo from such services as Flickr or Google Images. Naturally, the photo types derived this way may be somewhat ambiguous. Indeed, a wedding photo may well be a family photo.

To simplify matters, we desire a dataset in which the photo types are as unambiguous as possible. In addition, we seek the types of photos that are common enough to be of sufficient interest as consumer photos. With these goals in mind, a few experimental inspections indicate that four categories of photos fit our objectives well. They are *family*, *group field trip*, *sports team*, and *friends outing*.

We collected 10K photos. Some are from [22] and some are from online image services with relevant keyword queries. A group of human subjects then pick the 1K most unambiguous, properly fit photos for the 4 categories (250 each). For examples of these photos, see Fig. 4.5.

4.3 Method

Our method works by measuring the spatial similarity of the facial arrangement and the attribute similarity of the faces of the photos. Specifically, let us first consider computing the similarity score of two photos.

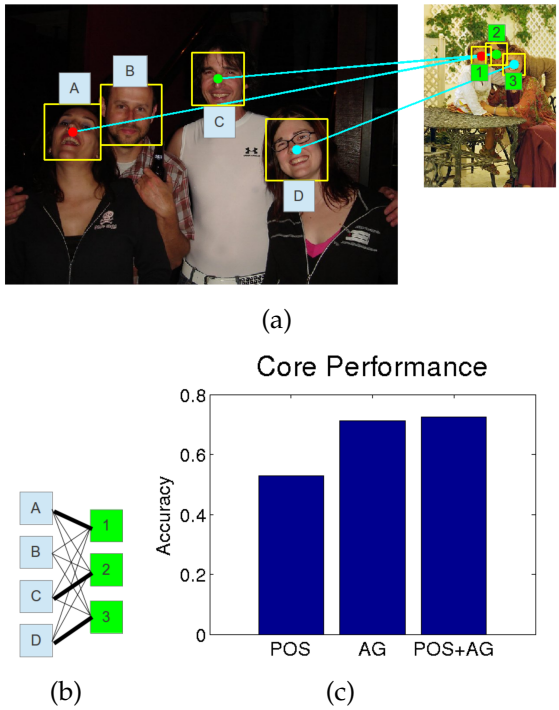


Figure 4.3: (a) and (b) are two sample photos showing the face bipartite graph. (c) is the core experiment result.

4.3.1 Bipartite Matching

For the two photos shown in Fig. 4.3(a), first detect the face bounding boxes. Then, we represent the faces as the nodes of a bipartite graph as shown in Fig. 4.3(b). Associated with each edge is a weight $w_{i,j}$ that captures the cost of matching the respective pair of faces, face i from the first photo and face j from the second, as a corresponding pair. Then, we find a maximum assignment (one in which the node set, say the right hand side in this example, with the smaller number of nodes has all its nodes matched) of minimum weight. Naturally, a matching must respect the one-to-one relationship. This is an example of the minimum weight bipartite assignment problem, which can be readily solved by the Hungarian algorithm [37, 48].

Edge weights are determined according to Eq 4.1 with the intent that a smaller

weight implies a higher degree of similarity. Here, the weight of edge (i, j) is a linear combination of the positional term and the attribute terms. The weights of the other edges are computed in a similar fashion.

$$w_{i,j} = \alpha||\mathbf{x}_i - \mathbf{x}_j|| + \sum_l \beta_l h_l(a_l(i), a_l(j)) \tag{4.1}$$

Positions. The faces coordinates of each image are first normalized so that the median face sizes in the two images are the same. Then, the faces coordinates within each photo are mean removed. The norm of the positional difference is weighted by α.

Attributes. Each face is associated with it a set of attributes indexed by l. For instance, $a_l(i)$ is the value of attribute l of face i. Function h_l computes the difference of two attribute l values. Each attribute difference is weighted by β_l.

Let us denote by w^* the sum of the weights of the matching edges selected. Fig. 4.3(b) shows the optimal set of edges selected using positions alone ($\alpha_i = 0$ for all i) as the darkened edges.

4.3.2 Face Number Discrepancy

Notice that this matching algorithm does not require the two photos to have the same number of faces. Such requirement would be too stringent. Firstly, the number of training data would greatly decrease when partitioned into photos of different numbers of faces. Secondly, as are evident in the examples we show, photos of different numbers of faces may still have structural similarity that is relevant. On the other hand, allowing face number discrepancy may be unfair in certain cases. Indeed, a two-person photo is very likely to match well with

a group photo simply due to chance. To address these issues, we pay an additional cost for any face number discrepancy and define the final similarity of two photos I_1 and I_2 as

$$d(I_1, I_2) \equiv w^*_{I_1,I_2} + \gamma |I_1 - I_2|_{\text{face}}, \tag{4.2}$$

where $w^*_{I_1,I_2}$ is the weight of the optimal matching of I_1 and I_2, and $|I_1 - I_2|_{\text{face}}$ is the difference of the numbers of faces in the photos.

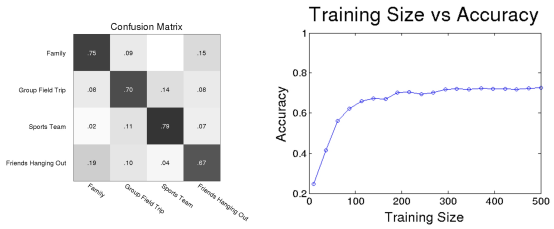
With the ability to compute the pairwise similarity score between any two photos, we can readily use many standard classifiers to complete the photo classification algorithm. For simplicity, we use k-NN through this work.

4.4 Experiments

Here, we describe the experiments we conduct to evaluate our method. We randomly split our dataset into 50% for training and 50% for test. The face attributes we use are age and gender. We use the predictions from the algorithm proposed by [11], in which there are 7 age bins roughly representing different stages of life. For h_{gender} we use the binary gender difference, and for h_{age} we use the bin difference which roughly captures how far apart two ages are. Finally, we use leave-one-out cross validation on the training set to tune the parameters required in our algorithm.

4.4.1 Main

The core performance results are summarized in Fig. 4.3(c). We carry out three sets of experiments. For the positions only experiment, which we denote by



(a) (b)

Figure 4.4:

POS, we turn off the face attributes ($\beta_{\text{gender}} = \beta_{\text{age}} = 0$). For the age and gender experiment, denoted by AG, we turn off the position contribution ($\alpha = 0$). When combining positions with age and gender, which we denote by POS+AG, we attain an accuracy of 72.6%. Fig. 4.4(a) shows the confusion matrix of the 72.6%-accuracy experiment. Compared to the 52.8% accuracy of POS, AG achieves an accuracy of 71.2%. It is perhaps not too surprising that age and gender attributes seem to play an important role in photo classification. From human intuition there is no shortage of plausible reasons behind it. Indeed, a family photo tends to include a wider range of faces of disparate age ranges and genders. On the other hand, a photo of a group of friends hanging out tends to have most of the faces belonging to roughly the same age groups, and they tend to be either all males or all females. Likewise, for a sports team photo, in most of the cases it consists of a majority of all males or all females, as official sport teams are rarely coed. For a group photo of field trip, the faces usually contain a more even mix of males and females, and the age range is wider as well.

Positional cues have their own merits, however. The performance of POS at 52.8% outperforms the random guess accuracy of $1/4 = 25.0\%$. Considering that no appearance-based cues are used in POS, this result quantitatively supports our hypothesis that facial position arrangement gives a nontrivial cue that can be helpful for photo type classification. In addition, POS+AG does give a

significant, albeit small, improvement of 1.4%.

We also use purely appearance-based features as a rudimentary baseline for our task of photo type classification. The feature we use are GIST [1], with RBF SVM as the classifier whose parameters we tune by cross validation on the training set. The performance result is a mediocre 42.3%, compared apple-to-apple with those results shown in Fig. 4.3(c). We do not find this result surprising. After all, a dominating factor for determining the type of a consumer photo is the humans in the photo. As such, a method that directly analyzes the humans in the photo may likely work better than one that does not.

4.4.2 Horizontal Symmetry

It is interesting to point out that we can effectively double the number of training data for POS by symmetrically flipping each training photo left-to-right. This observatoin comes from the assumption that, everything else being equal, people have no preference for the left or right side in a photo shot. Indeed, allowing such symmetry turns out to improve the performance of POS by 2-3%. Throughout this work, the experiments we conduct for POS and POS+AG take advantage of this horizontal symmetry assumption.

4.4.3 Effect of Training Size

While fixing the same test set, we artificially change the size of the training set down to as few as 10 photos. Fig. 4.4(b) gives the result for both POS and AG. In both cases, we see that the accuracy of > 50% for POS and that of > 70% for AG

are both achieved in as few as 100 and 200 training photos, respectively. In general, we find that contextual cues usually require much less training data than appearance-based features to attain their optimal classification performance.

4.5 Conclusion

In this work, we demonstrate the usefulness of facial arrangement and attribute (age and gender) cues in photo classification. Of course, there are limitations. For example, we may wonder how likely a truly randomly chosen consumer photo from the internet will be, say, a field trip photo given that it is quintessentially similar to the field trip photos in the training set. Nevertheless, the performance results from our work confirms the benefits of such contextual cues and encourages future work to build on it.

Figure 4.5: Shown in each of (a) - (h) are two image pairs. In each pair, the left image is the test query and the right is its most similar image from the training set. The left pair of images is based on POS, and the right pair of images is based on POS+AG, in which the gender and age predictions are shown as well. The ground truth photo types are provided at the bottom of each image. Best viewed in magnification in color.

BIBLIOGRAPHY

[1] A. Oliva abd A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. In *Proc. IJCV*, 2011.

[2] M. Abdel-Mottaleb and L. Chen. Content-based photo album management using faces' arrangement. In *Proc. ICME*, 2004.

[3] M. Abdel-Mottaleb and L. Chen. Content-based photo album management using faces arrangemen. In *Proc. ICME*, 2004.

[4] S. Ali, O. Javed, N. Haering, and T. Kanade. Interactive retrieval of targets for wide area surveillance. In *Proc. ACM*, 2010.

[5] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. In *Proc. IJCV*, 2011.

[6] A. Barbu. Learning real-time mrf inference for image denoising. In *Proc. CVPR*, 2009.

[7] D. Batra, A. C. Gallagher, D. Parikh, and T. Chen. Beyond trees: Mrf inference via outer-planar decomposition. In *Proc. CVPR*, 2010.

[8] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In *Proc. CVPR*, 2004.

[9] M. Blaschko and C. Lampert. Object localization with global and local context kernels. In *Proc. BMVC*, 2009.

[10] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *Proc. ICCV*, 2001.

[11] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *Proc. CVPR*, 2013.

[12] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. In *Proc. IEEE*, 2001.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[14] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Proc. ECCV*, 2010.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[16] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *Proc. CVPR*, 2009.

[17] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. In *Proc. PAMI*, 2010.

[18] A. Gallagher and T. Chen. Estimating age, gender and identity using first name priors. In *Proc. CVPR*, 2008.

[19] A. Gallagher and T. Chen. Finding rows of people in group images. In *Proc. ICME*, 2009.

[20] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.

[21] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.

[22] A. Gallagher and T. Chen. Understanding images of groups of people. In *Proc. CVPR*, 2009.

[23] A. C. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *Proc. CVPR SLAM*, 2007.

[24] C. Galleguillos, B. McFee, S. Belongie, and G. Lanckriet. Multi-class object localization by combining local contextual interactions. In *Proc. CVPR*, 2010.

[25] X. Geng, Z.-H. Zhou, Y. Zhang, G. Li, and H. Dai. Learning from facial aging patterns for automatic age estimation. In *ACM MULTIMEDIA*, 2009.

[26] G. Guo, Y. Fu, C. Dyer, and T. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. In *IEEE Trans. on Image Proc.*, 2008.

[27] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: using appearance models and context based on room geometry. In *Proc. ECCV*, 2010.

[28] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *Proc. ECCV*, 2008.

[29] D. Hoiem, A. Efros, and M. Hebert. Puting objects in perspective. In *Proc. CVPR*, 2006.

[30] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *Int. J. Comput. Vision*, 75(1), 2007.

[31] D. Hoiem, A. A. Efros, and M. Herbert. Recovering occlusion boundaries from an image. In *Proc. IJCV*, 2011.

[32] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking in multiple cameras with disjoint views. In *Proc. ICCV*, 2003.

[33] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras. In *Proc. CVPRV*, 2005.

[34] W. Jia, X. He, H. Zhang, and Q. Wu. Combining edge and colour information for number plate detection. In *Proc. IVCNZ*, 2007.

[35] Z. J. Jia, A. Gallagher, Y.-J. Chang, and T. Chen. A learning based framework for depth ordering. In *Proc. CVPR*, 2012.

[36] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance. In *Proc. CVPR*, 1999.

[37] H. W. Kuhn. The hungarian method for the assignment problem. In *Naval Research Logistics Quarterly*, 1955.

[38] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *Proc. ICCV*, 2009.

[39] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. In *Proc. PAMI*, 2011.

[40] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Describable visual attributes for face verification and image search. In *Proc. PAMI*, 2011.

[41] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proc. ICCV*, 2005.

[42] J. E. Lee, R. Jin, and A. K. Jain. Rank-based distance metric learning: An application to image retrieval. In *Proc. CVPR*, 2008.

[43] Y. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *Proc. CVPR*, 2010.

[44] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *Proc. CVPR*, 2009.

[45] C. Li, D. Parikh, and T. Chen. Exploiting regions void of labels to extract adaptive contextual cues. In *Proc. ICCV*, 2011.

[46] J. Lim, P. Arbel andez, C. Gu, and J. Malik. Context by region ancestry. In *Proc. ICCV*, 2009.

[47] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *Proc. CVPR*, 2010.

[48] Alexander Melin. http://www.mathworks.com/matlabcentral/ fileexchange/11609.

[49] A. C. Murillo, I. S. Kwak, L. Bourdev, D. Kriegman, and S. Belongie. Urban tribes: Analyzing group photos from a social perspective. In *Proc. CVPR*, 2012.

[50] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. *JCDL*, 2005.

[51] M. Nikolova. Model distortions in bayesian map reconstruction. In *AIMS J. on Inverse Problems and Imaging*, 2007.

[52] D. Parikh, C. Zitnick, and T. Chen. From appearance to context-based recognition: Dense labeling in small images. In *Proc. CVPR*, 2008.

[53] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. In *IEEE TIP*, 2003.

[54] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proc. ICCV*, 2007.

[55] C. Desai D. Ramanan and C. Fowlkes. Discriminative models for multi-class object layout. In *Proc. ICCV*, 2009.

[56] T.M. Rath and R. Manmatha. Word image matching using dynamic time warping. In *Proc. CVPR*, 2003.

[57] C. Rother, V. Kolmogorov, T. Minka, and A. Blake. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs. In *Proc. CVPR*, 2006.

[58] Bryan C: Russell and Antonio Torralba. Building a database of 3d scenes from user annotations, 2011.

[59] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *Proc. CVPR*, 2011.

[60] A. Saxena, M. Sun, and A. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. PAMI*, 31(5), 2009.

[61] A. Saxena, M. Sun, and A. Y. Ng. Make3d: Learning 3d scene structure from a single still image. In *Proc. PAMI*, 2009.

[62] H. Scharr, M. J. Black, and H. W. Haussecker. Image statistics and anisotropic diffusion. In *Proc. ICCV*, 2003.

[63] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proc. IJCV*, 2002.

[64] Mark Schmidt.

[65] Mark Schmidt.

[66] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on mrfs in low-level vision. In *Proc. CVPR*, 2010.

[67] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. CVPR*, 2006.

[68] P. Singla, H. Kautz, J. Luo, and A. Gallagher. Discovery of social relationships in consumer photo collections using markov logic. In *Proc. CVPRW*, 2008.

[69] J. Sivic, C. Zitnick, and R. Szeliski. Finding people in repeated shots of the same scene. In *Proc. BMVC*, 2006.

[70] Z. Song, M. Wang, X.S. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *Proc. ICCV*, 2011.

[71] Z. Stone, T. Zickler, and T. Darrel. Autotagging facebook: Social network context improves photo annotation. In *Proc. CVPR*, 2008.

[72] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. In *Proc. ECCV*, 2010.

[73] A. Torralba. Contextual priming for object detection. In *Proc. IJCV*, 2003.

[74] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.

[75] S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. In *Proc. ICCV*, 2009.

[76] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *Proc. ECCV*, 2010.

[77] G. Wang, A. C. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *Proc. ECCV*, 2010.

[78] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *Proc. CVPR*, 2011.

[79] Y. Weiss and W. T. Freeman. What makes a good model of natural images. In *Proc. CVPR*, 2007.

[80] X. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo.

[81] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *Proc. CVPR*, 2012.

[82] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proc. CVPR*, 2010.

[83] X. Zhu and D. Ramanan. ace detection, pose estimation and landmark localization in the wild. In *Proc. CVPR*, 2012.